图书情报工作
第58卷 第22期 2014年11月

# 国际重要科研机构网络信息采集与存档管理平台构建

吴振新 谢靖 张智雄 胡吉颖 李文燕

[摘 要] 介绍中国科学院文献情报中心正在建设中的国际重要科技机构网络信息采集与存档管理平台的整体框架,详细分析其技术平台架构组成和功能以及平台对IIPC基本工具的选择和使用情况,介绍所制定的资源采集政策、提供的开放接口以及存档平台的总体进展情况。

[关键词] 国际科研机构 网络信息 网络采集 网络保存
[分类号] G252 G259
DOI: 10.13266/j.issn.0252-3116.2014.22.016

## 1 前 言

随着网络技术的不断发展,越来越多的科技信息资源被发布到网络上,除了各种科研机构、研究团体、项目网站上的重要科技信息和成果外,还存在大量的科技新闻组、重要科学家个人网页、科学博客、科学论坛、开放原始科学数据、开放科研笔记本、大众科学的收藏等众多有价值的网络资源。这些科研信息对于科技决策人员、科技管理人员、科技情报人员及科研人员有着重要的价值。

网络存档是对Web上的信息资源进行收集、保存并确保这些资源能够被长期使用的一系列持续活动,为持久、有效地保存互联网资源提供了可靠的途径。2012年11月,美国国家数字信息基础设施保存计划(National Digital Information Infrastructure and Preservation Program,NDIIPP)发布了《处于危险中的科学:构建在线科学内容保存的国家战略》报告[1],在该报告中,明确了将在线科学内容保存提升成为美国国家战略。

由此可见,国际重要科技机构网络信息存档建设已经成为一项迫在眉睫的资源建设和保障任务,成为科技管理部门、科技文献情报机构迫切需要的资源保障和支撑系统。本文主要介绍中国科学院文献情报中心(以下简称"文献中心")正在开展的国际重要科技机构网络信息采集与存档管理平台的建设情况。

## 2 网络信息采集与存档管理平台的设计

### 2.1 设计思路与目标

在前期已经完成的国家社会科学基金项目"网络信息资源保存的理论与方法研究"的基础之上,笔者对现有的国际Web存档技术和工具进行了深入调研和实验,结合中心Web存档的实际需求,选用成熟的开源技术和工具,并对选用的开源工具的功能进行适当调整,通过一定量的个性化定制开发以满足个性化需求。在此基础上,整合构建成一个模块化的、开放架构、易于扩展升级的Web Archive采集与管理平台(NSL-WebArchive),它能够提供方便的个性化采集策略定制、自动调度的网页采集、半自动的采集资源初步加工、存档管理等功能,提供基于URL和内容的检索与获取服务。

### 2.2 总体架构

NSL-WebArchive是一个包括采集、管理、服务在内的三层架构(见图1)。在底层,通过部署多个采集结点来完成具体的网络资源采集工作。中间的Web Archive采集与管理平台,则负责对底层的多个采集结点进行配置、调度、管理,实现网络信息资源采集存档。上层的Web存档访问服务平台,为用户提供存档资源的访问和获取服务,同一层面的开放接口则是为其他的服务系统提供检索浏览及数据输出服务,使得用户可以从多种途径访问存档资源。

### 2.3 关键技术分析与选择

国际互联网保存联盟(International Internet Preser-

[作者简介] 吴振新,中国科学院文献情报中心研究馆员,E-mail: wuzx@mail.las.ac.cn;谢靖,中国科学院文献情报中心馆员;张智雄,中国科学院文献情报中心研究馆员;胡吉颖,中国科学院文献情报中心馆员;李文燕,中国科学院文献情报中心、中国科学院大学硕士研究生。
收稿日期:2014-09-03 修回日期:2014-10-23 本文起止页码:100-104 本文责任编辑:徐健

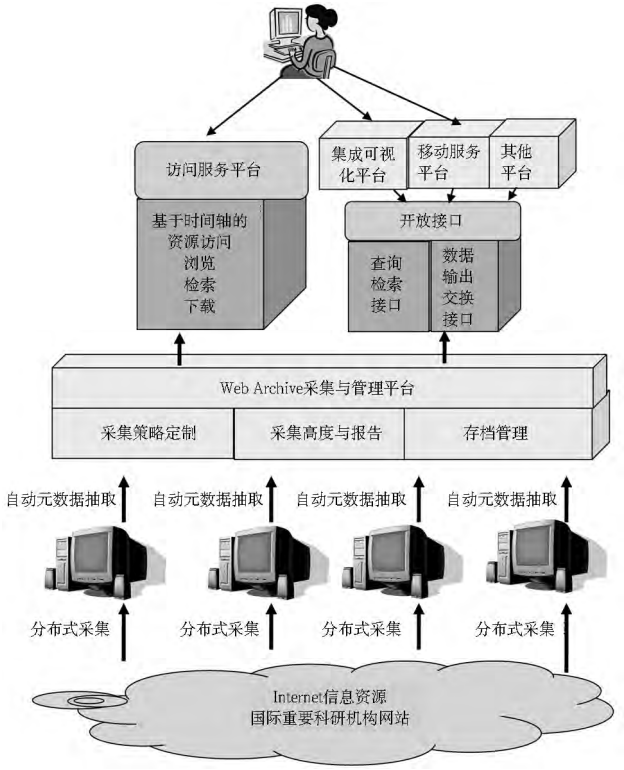


图1 NSL-WebArchive 总体建设架构

vation Consortium，IIPC)[2]资助开发了很多开源工具[3]，在全球得到了广泛部署和使用，国内也有文章进行了相关工具介绍[4]。笔者在考虑技术解决方案时，将IIPC的开源软件作为构建存档平台的首选，将相关工具作为框架中的组件纳入，不改变开源工具本身的功能，这样既可快速实现采集系统平台，还能够充分利用原有工具的优点，同时具有更好的兼容性，可实现无缝升级，尽享开源工具的优势。

2.3.1 采集技术与工具 IIPC提供的与采集和获取相关的工具包括ArchiveFacebook、Heritrix、HTTrack、SiteStory、WARCreate、Warrick、Wget等，其中尤以Heritrix[5]应用最为广泛。

从功能角度看，Heritrix支持复杂的爬行定义和过滤，具有丰富的可配置功能，如抓取频率可设置为每天、每周、每月、每季度、每年。它采用广度优先算法，用来抓取完整的、精确的、站点内容的深度复制，重新抓取相同的URL时不删除原先的版本，同时保存多个版本。Heritrix非常适合大规模网络存档。

从开发角度看，它采用了模块化的设计，用户可以在运行时选择适用的模块。Heritrix由核心类(core classes)和插件模块(pluggable modules)构成。核心类可以配置，但不能被覆盖；插件模块可以由第三方模块取代，所以用户可以用第三方模块来取代默认的插件模块，从而满足自己的个性化抓取需要，很方便地实现高度可扩展性，这也是其最出色之处。

2.3.2 采集管理工具 IIPC提供的采集管理相关的工具包括CINCH、NetarchiveSuite、Web Curator Tool(WCT)，其中后两者都是底层调用Heritrix进行资源采集的。

NetarchiveSuite[6]是一个允许图书管理员自己定义和控制网页资料收割的工具。可以进行3种不同类型的Web采集：主题采集、选择性采集和整个国家域的快照，对非技术人员友好，具有低维护、高bit保存安全和松耦合的特点，由于用户相对较少，软件升级更新比较慢。WCT[7]主要作为选择性网络采集的管理工具，应用在图书馆和其他收藏机构，提供友好的Web管理页面，支持非技术人员进行Web资源采集和管理工作，其优势在于对网络采集过程的完全控制，目前由英国国家图书馆和新西兰国家图书馆负责升级维护。从NSL－WebArchive面向国际重要科研机构的采集目标来看，WCT是更为合适的选择。

2.3.3 访问和检索工具 访问和检索工具目前有Memento Time Travel、NutchWAX、WERA、Wayback Machine、Xinq，NSL-WebArchive选择了使用较为广泛的Wayback Machine和NutchWAX。

Wayback Machine[8]是目前Web Archive领域中广为使用的存档资源访问系统，它集索引、检索、再现等功能于一体，能够自动监测指定的目录实现WARC文档的增量索引，能够为用户提供基于URL的检索以访问Web Archive资源。

NutchWAX[9]即“Nutch(W)eb(A)rchive e(X)tensions”，则是目前Web Archive中应用最为广泛的全文索引开源软件工具，可以和Wayback配合为存档的Web数据提供索引和搜索的功能，它采用Lucene技术在Haphoop环境中构建，可以实现大规模数据的索引。

其他如WERA(WEb aRchive Access)，也提供类似Wayback Machine的访问功能，还可以提供全文本搜索以及不同版本网页的导航。Memento Time Travel则是谷歌浏览器的扩展；Xinq(XML INQuire)是一个XML Archive结构化存档内容的访问工具。

2.3.4 存储与维护管理工具 IIPC还提供了一系列开源工具用于集合存储和管理，如HTTrack2ARC、Java Web Arrchive Toolkit(JWAT)、JHOVE2、Web Archive Transformation(WAT) Format、Web Archive Transformation(WAT) Utilities、WarcManager、WARC Tools等，主要是对存档数据文件进行格式转换、内容抽取、内容识别验证的工具。其中WarcManager[10]是一个轻量级的

数据库 Web 应用程序，它为 WARC 数据集合提供了很好的浏览接口，提供探索 WARC 文件内容的功能，可以帮助快速浏览、检索和分析网络抓取数据的归档。NSL－WebArchive 在存储层选择了 WarcManager，用于从 WARC 文件中提取结构化的元数据，以便在后期开展基于 WARC 格式的数据内容分析。

### 2.4 技术框架与功能模块

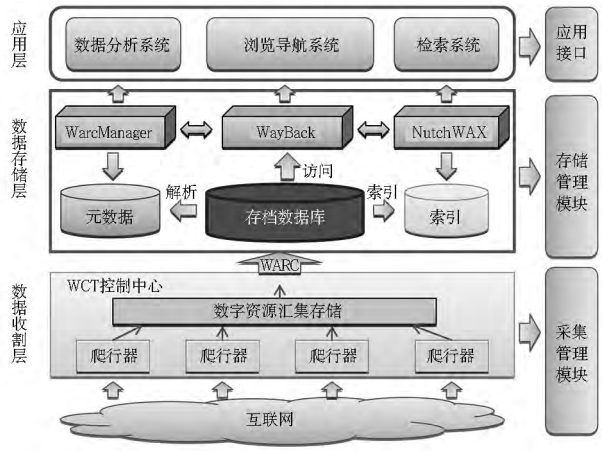


图 2 Web 采集与存档系统技术框架

技术架构采用 3 层结构，分为数据收割层、数据存储层、应用层。数据的收割和存储遵循 IIPC 框架结构。

2.4.1 数据采集层与采集管理模块 采用 IIPC 建议的 WCT 采集管理软件工具对采集任务、流程进行定制和规范管理。WCT 内嵌了开源采集器 Heritrix，采用分布式结构采集系统，将从 Internet 采集到的 Web 信息以国际标准 WARC 格式存储。

NSL-WebArchive 需要在采集层结合 WCT 构建采集管理模块，实现对采集结点的采集站点配置、采集任务调度等管理功能。

2.4.2 存储管理层与存储管理模块 在采集层，平台使用 WayBack 对存储的 WARC 数据建立基于 URL 的索引；使用 Nutch WAX 对 WARC 数据建立全文索引，提供检索查询功能；尝试使用数据分析抽取工具，将 Web 数据转换为半结构化的元数据，并对文本内容进行分析和挖掘，识别抽取知识对象存储在存档数据库中，通过扩展 WarcManager 对数据进行统计分析。存储管理系统对存档的 WARC 文档、存档数据库以及索引进行统一的管理和调度。

通过模块封装，平台将上述功能统一在存储管理模块进行调用。

2.4.3 应用层 应用层主要通过 WayBack 的访问和导航功能接口，为用户提供基于 URL 的检索、浏览功能，查看数据不同时间的发展变化。同时系统通过将 NutchWAX 引入 Wayback 系统的配置中，可以实现 Web Archive 的全文检索，即在 Wayback 和 NutchWAX 中进行一定的设置就可以将二者结合在一起，让 Wayback 使用 NutchWAX 索引，而 NutchWAX 的搜索结果也可以通过 Wayback 显示出来。同时，平台通过数据分析工具提供的多种接口，可以为系统提供数据的多角度分析揭示。

## 3 网络信息采集与存档管理平台的实现

### 3.1 目标资源筛选和存档策略制定

根据文献中心科技机构自动监测服务系统所积累的多家国际重要科研机构和科研管理机构信息以及其下属数百个重要研究院所及分支机构的信息，结合中国科学院的重点研究方向，筛选出美国、日本、韩国、法国、德国、俄罗斯、加拿大、澳大利亚、印度等国家数百个重要国立研究机构和研究管理机构，作为 Web 存档的采集目标资源。

采集目标确定后，还需要确定制定各机构网站的采集策略，NSL－WebArchive 制定了采集策略所要遵循的几项基本原则：

(1) 按照采集范围的不同，通常把采集策略分为完整性采集、选择性采集和混合型采集 3 种策略。完整性采集一般是对特定网络域的自动化的全面采集；把基于主题的以及其他的选择性策略合并为一种，统称为选择性采集策略；从项目层面，本项目所采用的是选择性采集，需要对重点领域筛选目标种子站点集合。

(2) 在采集策略的制定中，首先需要解决的问题是域的确定，通常采用多种方式确定网络域，对国家域的采集可以通过国家顶级域名、服务器的物理位置、网站所有者的国籍、网站内容相关性以及网站语言等多种方式确定。本项目中域的确定主要基于所提供的种子站点的域名，该域名下的所有页面以及页面中嵌入的链接，只要是基于该域名下的页面和内容(包括图片、视频等多媒体文档) 都进行采集存档，其他则不予存档。由于是 Archive 性质，每次采集都实施基于域名的完整性采集，并不考虑搜索引擎采集中的重复性过滤和变化比较等操作。

(3) 考虑到完整性采集的范围广、内容多、一次采集所需的时间较长，采集和存储成本都很高，因此选择合适的采集频率也是非常重要的，关于采集频率的确定，主要取决于采集预算和网络存档的预期目标。由于项目的采集目标是科技机构网站，资源变化不是很

频繁，而且目标种子数量也仅在数千个，所以目前的频率预定在每年12次采集。

(4) 网络礼仪的遵守。在网络信息采集过程中，一般情况需要遵守国际惯例的网络礼仪，对目标站点拒绝采集的内容予以回避。而本项目是出于全面采集资源进行长期保存，所以可以考虑跳过Robot协议，但采集速度要适中，避免对目标站点造成网络拥堵。

在实际存档中有些网站的采集策略需要调整，因此需要存档人员参与分析其网站结构和更新变化规律，制定适合各机构网站的采集策略和更新频率，并在平台上随时调整这些网站的采集频率。

### 3.2 构建 Web Archive 访问平台

对于已经采集存档的资源，需要构建Web Archive访问平台，为用户提供多种获取存档资源的途径。

作者设计了3种不同的访问服务：①最基本的是通过开源系统WayBack对存储的WARC数据提供基于URL的查询访问，为用户提供基于时间轴的网页访问功能，实现“时光回溯机”的功能。②采用Nutch-WAX对存储的WARC数据建立全文索引进行访问，为用户提供全文检索查询功能；③在完成上述功能的基础上，将探索使用数据分析抽取工具，将Web数据转换为半结构化的元数据，并对文本内容进行分析和挖掘，识别、抽取并存储知识对象，通过扩展WarcManager对数据进行统计分析。

### 3.3 构建开放服务接口

NSL-WebArchive还提供了可嵌入第三方服务系统的查询和检索接口，扩大存档的Web资源的使用范围。基于Web服务体系架构，让存档的Web资源可以被授权的第三方机构检索和利用。同时，查询和检索接口还支持不同的平台，可以嵌入到移动平台的文献服务系统。

开放的查询和检索接口支持基本的检索功能，包括：按指定元数据的检索、取回结果列表或单条详细信息。以Web Service的方式提供检索接口，分别利用基于SOAP协议(基于XML的标准API，以XML格式发送请求和返回结果)和REST协议(以URL格式发出请求，结果以XML格式返回)的Web Service实现平台检索功能的封装。

同时，实现存档资源的输出接口，遵循已定义的访问和数据共享规范，按定义的标准格式，自动(或半自动)地将数据转出，与其他系统实现共享。

### 3.4 进展与效果

目前NSL－WebArchive平台已经完成了初步建设工作，对优选的数百家主要科研机构和科研管理机构的相关网站，按照已经制定的采集策略基本原则分别进行深度分析，制定个性化的采集策略，并在Web Archive采集与管理平台上进行统一配置、部署、调度，存档人员在终端以多个定向采集系统为网页爬行工具，进行分布式的信息采集，将采集到的重要科研机构的网站内容存储到本地文件系统，同时建立索引提供访问服务，如图3所示：


图3 NSL-WebArchive访问服务平台页面

随着平台建设的初步完成，NSL – WebArchive 开始实施面向大规模的科技网络信息资源采集存档，不断改善平台性能，丰富存档资源的访问服务功能，并尝试利用知识技术进行存档资源深层分析和挖掘，探索如何利用存档资源更好地为学术研究和情报人员提供服务。

## 4 结 语

构建国际重要科研机构的 Web 存档，对目标网站各个时期的信息进行保存和镜像，能够在网站信息发生变化和相应的信息不存在时，为用户提供基于时间线的网页重现功能，有效地保障信息的可用性。

同时通过对重要科研机构网站信息进行完整、持续的 Web 存档，支持用户在时间线上对相应的网络科技信息进行研究和分析，对历史数据进行挖掘和再利用，发现主题的演化规律，把握政策和技术变化的趋势，以能够对相应科技政策和技术的效果进行评估，从而对构建长期的科技战略决策、领域内的长期变化趋势进行分析、预测未来发展趋势等工作提供重要的支撑工作，这也是文献中心未来的一项重要研究领域。

作为国家级的保存机构，文献中心通过相关领域技术的研究、开发、部署和实践，在构建开放架构、可扩展的网络信息采集和管理平台过程中，探索解决保存中的各种技术问题，真正实现了重要网络科技信息资源的保存，初步完成了保存目标，使得这种非常重要的开放资源已经实实在在地成为了国家科技战略资源保障体系中的一个重要组成部分，为我国的科技、经济以及社会的发展发挥了重要的支撑和保障作用。

参考文献：


[1] Toward a national strategy for preserving online science [EB/OL]. [2012 – 12 – 30]. http://www.digitalpreservation.gov/meetings/documents/othermeetings/science – at – risk – NDIIPP – report – nov – 2012.pdf.

[2] IIPC [EB/OL]. [2014 – 08 – 05]. http://netpreserve.org/.

[3] Tools and Software [EB/OL]. [2014 – 08 – 05]. http://netpreserve.org/Web – archiving/tools – and – software.

[4] 刘兰，吴振新，向菁，等. 网络信息资源保存开源软件综述 [J]. 现代图书情报技术，2009，25(5)：11 – 17.

[5] Heritrix [EB/OL]. [2014 – 08 – 05]. https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[6] NetarchiveSuite [EB/OL]. [2014 – 08 – 05]. https://sbforge.org/display/NAS/Releases + and + downloads.

[7] WCT [EB/OL]. [2014 – 08 – 05]. http://Webcurator.sourceforge.net/.

[8] Wayback [EB/OL]. [2014 – 08 – 05]. http://netpreserve.org/netpreserve.org/tools/openWayback.

[9] NutchWAX [EB/OL]. [2014 – 08 – 05]. http://archive – access.sourceforge.net/projects/nutch/.

[10] WarcManage [EB/OL]. [2014 – 08 – 05]. https://wiki.umiacs.umd.edu/adapt/index.php/WarcManager.


## Construction of Web Information Acquisition and Archive Management Platform of the International Important Institutions


Wu Zhenxin[1] Xie Jing[1] Zhang Zhixiong[1] Hu Jiying[1] Li Wenyan[1,2]

[1]National Science Library, Chinese Academy of Sciences, Beijing 100190

[2]University of Chinese Academy of Sciences, Beijing 100049



[Abstract] This paper introduces the three – layer framework of the important international scientific institution's web information acquisition and archive management platform, which is developed by NSL. Then it analyzes its technology architecture and functions in detail, talks about the selection and application of IIPC open – source software, describes the acquisition policies and its OAIs, and reports its overall progress.



[Keywords] international research institution web information web harvest web archive