网络信息资源保存的编目方法与系统研究*

□ 孙敏杰 吴振新 孙志茹 / 中国科学院国家科学图书馆 北京 100190

摘要：为了将长期保存的网络信息资源提供给用户利用，保存机构需要对这些资源进行一定的组织与质量控制。文章介绍了网络资源保存编目研究的发展现状，介绍了目前几种常用的编目方法：延续传统编目方法、基于web2.0标签技术的信息组织方法、基于大规模Web archive自动编目方法。并对两个新型的编目系统进行了深入的剖析：一个是瑞士的电子资源长期保存工程e－Helvetica，它将图书馆编目系统与长期保存系统的摄入流程进行集成整合，利用编目系统对长期保存资源进行组织与控制；另外一种是新加坡的网页标注系统WAWI，借鉴web2.0标签技术为长期保存系统中的网络资源添加标签，实现对资源内容语义层面上的控制。希望能为国内网络信息资源保存的相关研究与实践提供一些参考。该文为2009年第七期“网络信息资源保存”专题文章之一。

关键词：网络信息资源，长期保存，编目，标签
DOI：10.3772/j.issn.1673—2286.2009.07.004

1 引言

网络信息资源保存（Web Archive，简称WA）是对目标领域内的网络信息资源进行收集、保存、提供访问服务的活动。为用户提供访问服务是WA的最终目标，因此除了资源的采集与保存工作之外，为了便于对网络存档的管理和利用，大多数的项目在网络资源存档后还要对采集到的资源做进一步的加工和整理，将其整合到原有的资源体系中提供后续的服务，即通常意义上的（元数据）编目工作。如澳大利亚的PANDORA项目[1]是由图书馆员对采集到的网络出版物进行编目，把编目数据加入到国家图书馆书目库中，供读者检索使用。

在资源采集阶段通常会利用自动化的工具抽取简单的元数据，如采集时间、从采集对象网页头标（header）字段中抽取的网页标题、上次修改时间等相关元数据，但如果要进行更加详细的描述，就需要专门的工具和系统以及图书馆编目人员的参与。然而网络信息资源具有数量巨大、质量参差不齐且无序化等特点，对这类资源进行组织和控制要耗费大量的人力物力。因此如何更好的利用图书馆已有信息组织经验和现代IT技术来组织所采集和保存的网络信息资源，成为保存领域一个非常有价值的研究课题。

2 网络信息资源保存编目研究的发展现状

WA开展十几年来，全球大大小小的保存项目上百个，其中有些项目对网络信息资源的编目进行了有益的研究和实践，不断地探索新的技术和更为有效的方法。

（1）传统编目方法的延续

作为图书馆的传统信息组织方法，编目工作自图书馆存在以来就是其主要业务工作，在这方面已经积累了相当成熟的经验。这些已有的资料编目方法为网络信息资源的保存和组织奠定了很好的工作基础。其中的一些方法和系统也被用于现在的WA资源登记中。例如PANDORA[2]为了很好地组织其所收集的Web

* 本文系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”（项目编号：06BTQ025）的研究成果之一。

资料，使之便于用户查找，将其存档的资料加上标题（Title）并用在线编目系统进行DC元数据编目后，加入到国家图书馆书目库及其合作者的书目库中，为用户提供访问服务。荷兰国家图书馆的e-Dpot系统[3]也采用了类似的方法。

然而，Web资源体量巨大，即使是基于主题的选择性采集，每个主题可能包含的网站资源数目也会以千、万计算，同时Web资源类型复杂，这就很难或是不可能使用传统的方法来编目。美国国会图书馆的Minerva项目[4]收集了2000年选举、“911”事件、2002年冬奥会的Web资料，每个主题集合都限定了资源采集范围。如果为每个网站编目（MARC），对于选举和冬奥会是可能的，该主题网站的数量大概是5,000个，约9千万Web文档；但是对于“911”事件和“9-11” Remembrance，该主题网站的数量约有32,000个，约3.3亿个Web页，对这些集合，仅有2,300个被挑选出进行编目。

同时Web资源是易变的，通常变化还很不明显，而传统的编目方法不具备很好的可持续性，因此采用传统的编目方法来组织Web资源有很大的局限性。虽然一些折中的办法建议通过采用一些元数据标准而继续延用传统编目方法，但考虑到实际需求，还是需要探索新方法来管理Web资源。

（2）基于web2.0标签技术的信息组织方法

相对于WA所拥有的复杂资源，上面提到的编目方式能为用户提供的WA内容信息就过于简单，因此不利于用户对WA的查找和使用。而语义网的出现（web 2.0）虽然增加了WA的难度，但也为WA的资源管理提供了许多新的思路。标签技术对WA编目工作就起到了很好的推动作用。如Technorati[5]、Flickr[6]和del.icio.us[7]上的标签，实际上是集合众人的力量为在线资料编目。Wayfinder[8]就是借鉴了这一方法的WA访问工具，用户可以通过Wayfinder界面为访问的对象加标签或注释，同时还可以浏览其他用户对这一对象的描述内容。

这种方式将编目过程的控制分散化，一方面提高了编目速度，另一方面也是对内容的更广泛、深入地挖掘。但是这种编目过程的分散也会引起人们对这种任意标注的资料的准确性的关注。

（3）基于大规模Web archive自动编目的探索

为了更好地解决海量数据的编目问题，有项目[9]提出一种基于大规模扩充的自动编目方式，即通过抽取技术从网页中自动抽取元数据，并以结构化形式存储，以此来实现对保存资源的编目。

网页中可供抽取的信息有两种：一类是描述元数据，如嵌入在HTML Web文档中的META标签里的描述元数据；另一类是起源元数据（provenance metadata），如在Web文档采集的过程中所收集的原始信息，它对于保障存档和存档认证以及存档访问都很有必要。其中，描述元数据的抽取可以采用基于位置的信息抽取方法[10]，即依据网页文档的内在结构特征来完成数据抽取，采集的html文档被送入html剖析器中，依据制定的数据抽取规则，剖析器建立一个反映html标签等级的剖析树。该方法有很高的抽取准确率，但当目标网页的结构发生改变时，该方法将不可行。对于起源元数据来说，还可采用基于本体论的信息抽取方法[11]，即通过一个适应性网络信息抽取系统获取元数据信息，它使用领域知识来描述数据，包括关系、词频、上下文关键词，但目前基于本体论的研究还待深入。相信随着信息抽取技术的不断成熟，元数据抽取技术将在WA信息组织方面发挥重要作用。

3 网络信息资源保存的编目系统研究

3.1 基于摄入流程的WA编目处理—e-Helvetica工程

瑞士国家图书馆（Swiss National Library，简称SNL）的e-Helvetica工程[12]主要承担瑞士电子出版物的收集、组织、保存与传播的任务，以使这些数字资源在将来能够被用户所访问和利用。该工程由针对不同类型电子资源进行长期保存的几个试验项目组成，这些项目的资源收集与保存均由SNL与其他图书馆合作完成（以下统称为合作馆）。本文以其中用于保存网络信息资源的Web Archive Switzerland试验项目为主进行分析。

（1）系统功能架构

e-Helvetica功能架构遵循OAIS参考模型[13]，包括六个功能模块：摄入（ingest）、数据管理（data management）、存储（archival storage）、保存规划（preservation planning）、系统管理（administration）、存取访问（access）。如图1所示。

摄入模块负责摄入和处理合作馆的数字对象及其元数据，直到该数字对象及其元数据被保存到存储模块中。

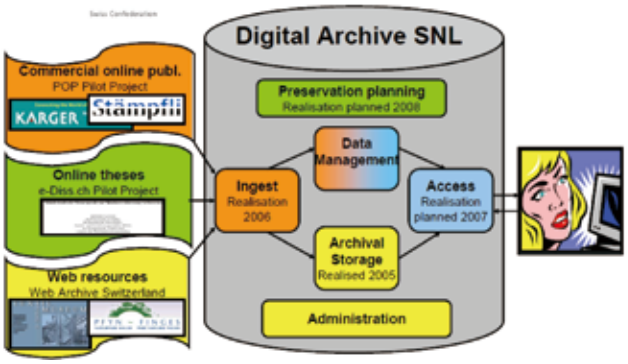

图1 e–Helvetica功能架构[14]

数据管理模块保存由摄入模块获取的所有元数据。

存储模块保存数字对象及其元数据的长期保存信息包。

保存规划模块负责制定OAIS模型的保存策略。

系统管理模块负责监控各个模块的运行。

存取访问模块，通过一个特殊的访问接口连接OPAC和SNL，使得用户通过图书馆OPAC访问包括长期保存资源在内的所有馆藏。

（2）数据处理流程

e-Helvetica对数字对象及其元数据的自动处理流

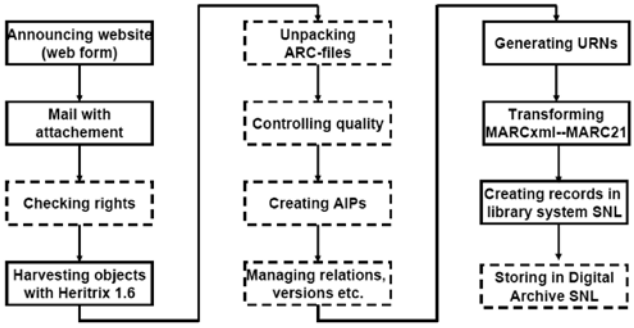

图2 e–Helvetica数据处理流程[15]

程，如图2所示。

资源登记：在Web Archive Switzerland试验项目中，合作馆使用web表单对收录的web站点资源进行声明（announcing website），即创建资源描述的元数据，该web表单是在都柏林元素集（DC）和MARCxml的基础上制定的。

元数据提交：元数据由分布在各地的合作馆创建，可采用两种提交方式：一种是电子邮件方式，合作馆将元数据包做为附件邮寄到SNL，适合小型图书馆；一种OAI-PMH方式，SNL直接从合作馆的存储系统里收割元数据，适合大型图书馆。这两种方式的元数据均以xml文件提交给e-Helvetica的摄入模块。

权限检查：检查收割方是否拥有长期保存和存取访问资源的权利。

数字对象收割：权限检查通过后，使用Heritrix收割数字对象。收割工具首先从之前提交的元数据中读取数字对象链接，然后根据该链接收割合作馆存储系统中的数字对象。

解压ARC文件包（unpacking ARC-files）：元数据和数字对象接收后，摄入系统会将其打包为ARC文件，在进行后续的处理之前需要解包。

质量检测：摄入过程的几个不同阶段都需要进行质量检查。包括数字对象病毒和完整性检测、元数据格式检测、重复提交控制机制。

创建AIP包（creating AIPs）：创建保存过程中的存储信息包。

关系、版本管理（managing relations, versions etc.）：对存档内容之间的关系及不同时期的存档版本进行管理。

生成URN：系统会为每个数字对象生成一个URN，该URN依据e-Helvetica长期保存资源唯一标识符即国家书目记录号（National Bibliography Numbers，URN:NBN）制定，并与URL相对应。

元数据完整性检测：摄入流程的最后一步是检测摄入阶段所收集的信息是否均已记录在元数据中。

数字对象及其元数据由摄入模块输出后，进入数据管理模块、存储模块和图书馆编目系统中，开始对资源进行管理、存储和组织。其中，完整的元数据包保存在数据管理模块；包含数字对象和元数据的长期保存信息包保存在存储模块；书目元数据格式由MARCxml转成MARC21后，被自动发送到图书馆编目系统，生成一条新的编目记录以供编目人员进行进一步的处理[16]。

由以上的数据处理流程，我们可以看到e-Helvetica工程基于长期保存系统的摄入流程集成了图书馆编目系统，用长期保存系统做前端的采集与后端存储，用图书馆编目系统对所采集的资源进行组织和质量控制。两者的结合，一方面使长期保存资源的质量得到保证，另一方面使图书馆编目系统突破传统资源的界

线，在网络资源组织方面得到了更好的传承。

3.2 基于web2.0技术的网页标注系统——WAWI

WAWI（Web Annotation for Web Intelligence）[17]是由新加坡南洋技术大学设计开发的网页标注系统，采用web2.0的标签方式实现，该系统与长期保存系统集成后，可供编目人员对WA资源进行著录或供用户为其添加标签。

（1）系统功能架构

WAWI系统作为 web curation流程的环节集成到WA长期保存平台中，集成后的系统架构如图3所示。

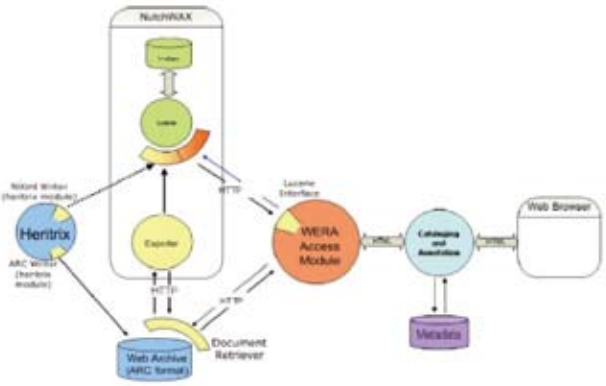

图3 WAWI标注与编目系统与IIPC Web Archive平台的集成[18]

该平台采用Heritrix采集web数据，采集下来的数据以ARC格式进行保存，NutchWAX对存档文件建立全文索引，然后WERA检索组件通过NutchWAX 对存档的纯文本和URL进行访问。WAWI通过WERA组件读取WA平台中的ARC文件，为编目人员提供网页标注服务，同时也为终端用户提供访问。

（2）WAWI系统的网页标注流程

整个系统流程分为三个阶段[19]：

第一阶段是标注表单（annotation schema）准备阶段。图书馆员使用标注表单管理器来创建元数据表单，该表单在浏览器中以树状视图的形式呈现，创建后被转换成xml文件保存到服务器端的数据库中。

元数据结构模型依据W3C语义网协会提出的Annotea系统，使用RDF框架。主要组成部分为：标注标题，预标注的目标文本，用户标注的标签或元数据，允许或访问的权限。此外，还包括唯一标识符、日期、url等信息。

第二阶段是元数据标注。系统在客户端浏览器中加载两部分内容：如图4所示，右侧以树状视图显示标注表单（元数据模型），左侧显示从WA存储系统（web archive repository）中读取的web页面。通过点击和拖拽动作，预标注的目标文本部分被高亮显示，同时被捕捉到标注表单中。用户完成标注后，表单将被保存在服务器端，以供以后的检索和审核。

图4 标注过程页面[20]

第三阶段是元数据检索阶段。标注阶段所创建的元数据及捕捉的目标文本均可提供检索。用户的检索指令被转换成XQuery查询发送到服务器端的XML数据库中，该数据库将结果返回到检索结果页面，该页面在显示标注元数据的同时，也提供到存档的web页面的链接。如果要同时检索元数据和存档的web页面，可以将WERA检索到的存档页面的纯文本和URL整合进WAWI元数据搜索引擎中来。

通过WAWI系统的标注流程，我们可以看到该系统的独到之处为采用了Context-sensitive标注方式，即在标注过程中建立元数据与存档的web页面的关系，这样编目人员就能够参照原文内容来著录元数据，该方式可以能够确保机构环境下WA编目过程的一致性和质量。

4 结语

可以看到，长期保存领域的编目研究与实践已经取得了一定的进展。在研究方面，学者们尝试着从不同的角度进行探讨：继承传统编目方法，尽可能的发挥图书馆在信息组织方面的长处；引入web2.0环境

下“全民织网”思想，采用标签来组织长期保存的资源，充分挖掘用户在资源组织方面的潜力；采用元数据自动抽取挖掘技术，减少了人力的投入，对长期保存领域的大规模采集具有重要的意义。在实践方面，除了本文所介绍的e-Helvetica、WAWI这两个典型系统外，还有诸多优秀系统，如澳大利亚国家图书馆的PANDORA系统、荷兰国家图书馆e-Depot系统等，这些系统对图书馆在长期保存领域的信息组织实践起到了很好的推动作用。

但同时我们也注意到，由于受到网络信息资源自身的特点以及人力、技术等诸多因素的限制，目前长期保存领域的信息组织方法和实践还处于不成熟的阶段，相关研究还有待于进一步的深入。随着网络信息资源数量的不断增大、信息类型日益复杂，网络信息资源保存的组织与控制将面临更大的挑战。

参考文献

[1] The PANDORA Digital Archiving System (PANDAS): Managing Web Archiving in Australia: A Case Study[EB/OL]. [2009-05-08].http://www.nla.gov.au/nla/staffpaper/2004/koerbin2.html.
[2] 同1.
[3] e-Depot and digital preservation[EB/OL].[2009-05-08].http://www.kb.nl/dnp/e-depot/e-depot-en.html.
[4] SCHNEIDER S M, et al. Building thematic web collections: challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive[C]// 3rd ECDL Workshop on Web Archives, Trondheim, Norway, August, 2003, 21.
[5] Technorati[OL].[2009-05-08].http://technorati.com/.
[6] Flickr[OL].[2009-05-08].http://www.flickr.com/.
[7] Delicious[OL].[2009-05-08].http://delicious.com/.
[8] DOUGHERTY M. Wayfinder: Building an interface for a Web archive[C]// 7th International Web Archiving Workshop. Vancouver, Canada, 2007.
[9] MASANÈS J. Web Archiving[M]. New York: Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.
[10] 龙丽, 庞弘燊. 国外web信息抽取研究综述[J]. 图书馆学刊,2008(5):13-16.
[11] 同10.
[12] e-Helvetica: Collecting and archiving digital publications[EB/OL]. [2009-05-08].http://www.nb.admin.ch/slb/slb_professionnel/01693/index.html?lang=en.
[13]邓君.基于OAIS与OAI-PMH的数字档案馆共享功能框架设计[J],档案学通讯,2008(3)
[14] Barbara Signori. web archive Switzerland[OL].
[15] 同14.
[16] Barbara Signori. e-Diss.ch: collecting and archiving online theses at the Swiss National Library[OL].
[17] WU P H J, HEOK A K H, TAMSIR I P. Annotating the Web Archives–An Exploration of Web Archives Cataloging and Semantic Web[J]. LECTURE NOTES IN COMPUTER SCIENCE, 2006,4312 :12.
[18] 同17.
[19] WU P H J, TAMSIR I P, HEOK K Y. Adrian. Applying Context-Sensitive Web Annotation in Evidence-based, Collaborative Web Archives Cataloging[C]//, International Workshop on Archiving Web, 2006.
[20] 同19.

作者简介

孙敏杰（1979-），中国科学院国家科学图书馆2008级硕士研究生，信息检索与技术方向。通讯地址：北京市北四环西路33号，中国科学院国家科学图书馆615，100190。E-mail：sunminjie@mail.las.ac.cn
吴振新，中国科学院国家科学图书馆副研究馆员，数字资源长期保存方向。通讯地址：北京市北四环西路33号，中国科学院国家科学图书馆，100190。E-mail: wuzx@mail.las.ac.cn
孙志茹，中国科学院国家科学图书馆2006级博士研究生。

Reaseach on Catalogue Method and System of Web Archive

Sun Minjie, Wu Zhenxin, Sun Zhiru / National Science Library, Beijing, 100190

Abstract: In order to make the archived web resources available to users, catalogue and quality control should be done by archive organizations. This paper introduces current researches on catalogue of web archive, and describes three methods: method of extending traditional cataloging , method based on web2.0 tagging technology, method based on the large-scale automatic cataloging. Finaly, it gives an in-depth analysis of two new cataloging system: one is a Swiss long-term preservation project e-Helvetica, it integrates the ingest process of long-term preservation system with the library catalog system, library catalog system is used on web source cataloging; the other is the Singapore’s page marked system WAWI, it applies web2.0 tagging technology to add labels for web resources. This paper is hoped to provide some references to native related research and practices.

Keywords: Web resources, Web archive, Catalogue, Tag

（收稿日期：2009-05-15；责任编辑：虞敏）