2014年资源联建项目培训

——网事典藏项目建设标准及方法

国家图书馆 数字资源部 推广工程培训与资源建设组

李云龙

2014年12月

主要内容

一 • 项目概况

二 • 工作流程

三 • 成果提交

四 • 与政府公开信息项目的区别

一、项目概况

1 项目背景

- 互联网上有大量有关人类社会、政治、历史、文化的信息，网络资源是一个国家文化遗产的重要组成部分，对于传承人类文化遗产和学术研究具有非常重要的作用，但由于互联网资源具有易失性，每天都有海量有价值的信息在消亡。

- 为了保存互联网上的信息，从20 世纪90年代中期开始，就有很多国家开展了网络资源存档项目，采集并保存了大量的互联网资源。

一、项目概况

IIPC简介

- IIPC（International Internet Preservation Consortium）
- 一个会员组织，成员超过25个国家，包括政府、大学、图书馆和档案馆。
- 致力于改善网页存档的工具、标准和方法，促进国际合作，收集、存储发布在互联网上的网页。
- 制定WARC存档标准，开发IIPC WARC分析工具。

网址 http://netpreserve.org/

一、项目概况

1 项目背景

- 我国也早已认识到互联网信息保存与保护的重要性，从 2002 年就开始了相关的实验和研究。
- 2009 年成立了国家图书馆互联网信息资源保存保护中心，专门致力于中国互联网信息资源采集与长期保存方面的研究工作。
- 为保存具有重要价值的互联网资源，推广工程将开展网络资源采集与长期保存工作。

一、项目概况

2 参与馆与数据量要求

- 先期参与试验的**5个**省级图书馆

首都图书馆

浙江图书馆

吉林省图书馆

湖北省图书馆

新疆生产建设兵团文化中心

- 每馆完成不少于**100个**网站的采集和长期保存

一、项目概况

3 建设内容

- 以**政府网站**的采集和存档为重点
- 主要采集反映地方政治、经济、文化发展等信息的**政府网站**或**重要的事业单位网站**（包括但不限于.gov.cn域名内的网站）
- 将采集到的网站进行**编目和发布**
- 采集**行政上从属于本地区**的政府和事业单位网站

二、工作流程

1. 采集准备
2. 资源采集
3. 加工编目
4. 数据预发布
5. 数据审校
6. 数据发布
7. 数据维护和长期保存

二、工作流程

1 采集准备

- 将需要采集的政府网站网址（URL地址）整理成采集列表（**excel表格**，表头如下），提交国家图书馆审核，审核通过作为正式采集列表进行采集。

政府网站采集列表

序号	网站名称	网站域名	备注

二、工作流程

2 资源采集

- 利用**网络采集软件**，对政府网站进行全面采集。
- 要求所采集的文件包含采集列表中政府网站域名内的全部内容，但**不包括**论坛等需链接**后台数据库**的内容。
- 所采集的文档格式遵循**WARC 1.0**标准，不含病毒、垃圾文件及采集列表外的其他信息。

二、工作流程

WARC 1.0标准

- ISO 28500：2009

- 存储主流互联网应用层协议；存储与其他已存数据相关联的任意元数据；支持数据压缩和维护数据记录的完整性；储存所有来自采集协议的控制信息；存储与其他存储数据的数据转换的结果；存储与其他存储的数据相关联的重复检测事件；在中断现有的功能前提下的扩展；支持在需要的位置通过截断或分割处理过于长的记录。

网址 http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717

二、工作流程

2 资源采集

- **每个网站单独采集**，采集结果以**文件夹**形式存储，文件夹名称以对应的**加工编号**来命名。（作为成果提交）

二、工作流程

加工编号命名规则

与下发文件有所不同，进行了修改

■ 政府网站采集加工编号命名规则（21位）

例如　首都图书馆采集北京市人民政府网站　GOV010001000120141115

GOV	机构代码	机构类型代码	流水号	采集日期
（3位）	（4位）	（2位）	（4位0001-9999）	（8位）

说明	
机构代码	各地方馆的机构代码，参照“地方文献数字化机构代码”
机构类型代码	政府机构（01） 事业单位（02）
流水号	顺序排列，每个网站需具有固定的流水号，不同时间多次采集同一个网站时，该网站的流水号不变。
采集日期	资源被采集的日期，格式为yyyymmdd（例：20141115）

二、工作流程

2 资源采集

- 对采集完成的数据要填写“**网络采集统计信息表**”（作为成果提交）

表一：网络采集统计信息表

加工编号	网站名称	网址	任务名	压缩前大小（GB）	压缩后大小（GB）	URL数量	运行时间	采集日期

二、工作流程

3 加工编目

- “政府及相关机构存档网站元数据著录规则”
- **每个采集结果**对应一条完整的编目信息
- 将编目信息制作成**excel表**（该表中**访问地址**一项需要在数据发布后才能取得补齐）（作为成果提交）
- 需要在唯一标识符系统中注册**CDOI**

二、工作流程

3 加工编目

“政府及相关机构存档网站元数据著录规则”

- 著录对象。著录对象为存档政府和事业单位网站（简称政府网站）。以具体的存档政府网站对象为一个著录单位。如果一个政府网站具有多个主页域名，著录时作为一个对象著录，多个域名作为子项著录。

- 著录要求。对采集到的网站进行编目加工，要求参照相关元数据规范和著录规则进行编目，且准确无误。编目数据以EXCEL文件形式提交。

二、工作流程

3 加工编目

政府及相关机构存档网站基本字段及著录要求

元素	修饰词	元素的著录内容	元素必备性、可重复性
标识符	加工编号	给予资源的一个明确标识，可以唯一标识元数据的标识符。具体规则见备注	必备，不可重复
	CDOI	唯一标识符号	必备，不可重复
网站名称	正题名	著录网站名称。信息取自网站页面首页源代码中的title。例如：“北京市朝阳区人民政府网站”的title是“北京朝阳”，将“北京朝阳”著录为网站名称。若title为空，或不反映网站内容，可用banner或其他明显反应网站内容的名称。	必备，不可重复
	交替名称	按照统一要求著录“**网站”作为交替名称，对网站名称进行解释说明。例如：朝阳区人民政府的网站统一著录交替名称为：“北京市朝阳区人民政府网站”。	必备，可重复

二、工作流程

3 加工编目

描述		对网站内容的任何说明	必备，可重复
类型		所采集的政府网站类型：政府机构网站、事业单位网站	必备，不可重复
语种		资源所使用的语种	必备，可重复
格式	文件格式	所采集的网站资源存档格式。在本规范中为：WARC	必备，不可重复
	文件数量	著录WARC文件的数量	必备，不可重复
	文件大小	著录所有WARC文件的总容量	必备，不可重复
机构	机构名称	著录网站的所属机构的名称。例如：北京市朝阳区人民政府网站的所属机构是：北京市朝阳区人民政府	必备，不可重复
	机构级别	著录网站的所属机构的级别：国家级、省部级、司局级、县处级、乡科级。例如：北京市朝阳区人民政府的机构级别是：司局级	必备，不可重复
	交替名称	著录除正式机构名称以外的翻译名称、缩写名称、统一名称、并列名称等。例如：中华人民共和国外交部 其交替名称为：外交部	有则必备，可重复

与下发文件有所不同，添加了字段

二、工作流程

3 加工编目

域名		著录网站的主页域名，如果一个网站有多个域名，则全部著录。多个域名重复本元素。	必备，可重复
采集地址		政府网站的网址	必备，不可重复
日期	采集日期	资源被采集的日期	必备，不可重复
访问地址		所采集的政府网站存档访问网址	必备，不可重复
权限		与本资源有关的权利和许可的声明	有则必备，可重复
	版权信息	资源所有权信息	有则必备，可重复
数据提交单位			必备，不可重复
数据提交日期			必备，不可重复

二、工作流程

4 数据预发布

- 将采集到的文档（WARC文档）数据**进行索引**后预发布
- 取得所采集政府网站的**访问地址**，在编目数据中补齐访问地址字段的内容。
- 发布方式由各馆根据实际情况自行选择

二、工作流程

4 数据预发布（效果）

二、工作流程

4 数据预发布（效果）

二、工作流程

5 数据审校

- 编目数据审校：对编目完整的数据按照“著录规则”进行审校，保证各字段的准确、完整。

- 对象数据审校：在**预发布页面**通过点击的方式进行查验，保证页面内容都能正常打开，且与原网站保持一致。**文件夹命名**无误，与编目数据一一对应。

- 数据本馆审校合格后交**第三方进行验收**，验收不合格需要修改或重采，直到验收合格。（验收报告作为成果提交）

二、工作流程

6 数据发布

- 验收合格后，将最终成果进行正式发布，并提供用户服务。
- 同时将成果提交至国家图书馆。

二、工作流程

6 数据发布（效果）

http://navi.nlc.gov.cn/

部委名称	抓取时间
中华人民共和国外交部	2014年1月9日
中华人民共和国国防部	2014年1月9日
中华人民共和国国家发展和改革委员会	2014年1月10日
中华人民共和国教育部	2014年1月13日
中华人民共和国国家民族事务委员会	2014年1月14日
中华人民共和国公安部	2014年1月14日
中华人民共和国监察部	2014年1月14日
中华人民共和国民政部	2014年1月14日
中华人民共和国司法部	2014年1月16日
中华人民共和国人力资源和社会保障部	2014年2月10日
中华人民共和国国土资源部	2014年2月12日
中华人民共和国环境保护部	2014年2月11日
中华人民共和国交通运输部	2014年2月11日
中华人民共和国水利部	2014年2月11日
中华人民共和国农业部	2014年2月11日
中华人民共和国文化部	2014年2月17日

二、工作流程

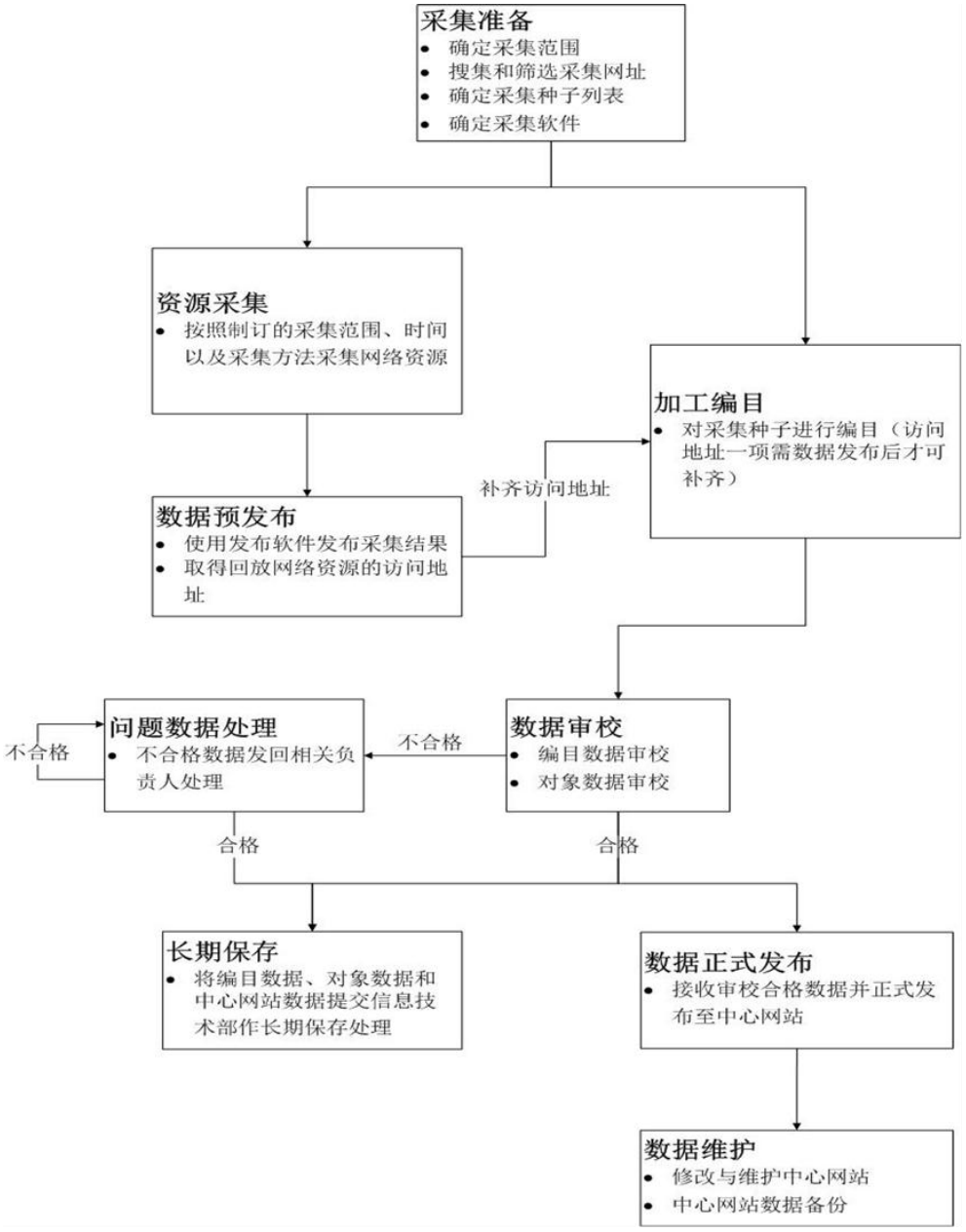
6 数据发布（效果）

二、工作流程

7 数据维护和长期保存

- 各馆负责对本机构制作及发布的信息及其发布网站进行长期维护，保障数据准确无误，显示正常，同时做好数据备份与长期保存工作。

二、工作流程


采集准备
确定采集范围
搜集和筛选采集网址
确定采集种子列表
确定采集软件
资源采集
按照制订的采集范围、时间以及采集方法采集网络资源
加工编目
对采集种子进行编目（访问地址一项需数据发布后才可补齐）
补齐访问地址
数据预发布
使用发布软件发布采集结果
取得回放网络资源的访问地址
问题数据处理
不合格数据发回相关负责人处理
不合格
不合格
数据审校
编目数据审校
对象数据审校
合格
合格
长期保存
将编目数据、对象数据和中心网站数据提交信息技术部作长期保存处理
数据正式发布
接收审校合格数据并正式发布至中心网站
数据维护
修改与维护中心网站
中心网站数据备份

三、成果提交

- 编目数据：政府网站的编目数据以excel表格方式提交。
- 对象数据：政府网站的对象数据以文件夹（以加工编号命名）的形式提交网络采集软件所采集的**全部采集结果**。
- 统计信息表：网络采集统计信息表，详见表一。
- 第三方质检报告

注：以移动硬盘（数据迁移后硬盘返还提交馆）或其他方式进行提交

四、与政府公开信息项目的区别

- 主要体现在以下两个方面

采集和著录的对象不同

政府公开信息项目	网事典藏项目
网站中每一条具体的政府信息	整个政府网站

项目采集的目的不同

政府公开信息项目	网事典藏项目
是为了获取网页中想要的内容为主，对内容进行整合，以应用服务为主。	保存网页原貌，以存档为主。同一网站多次采集。

演示

- 网络采集软件Heritrix实际操作演示

谢谢！

联建网络资源采集QQ 群：

365776635