扬州市图书馆网事典藏建设分享

汇报人：朱静

目录

01 项目相关背景及建设指导思想

网事典藏项目相关背景及建设指导思想

网事典藏是什么？

网事典藏是一个将政府网站的全部网页、图片、资源文件等用WARC格式和纯文本保存存档的项目。

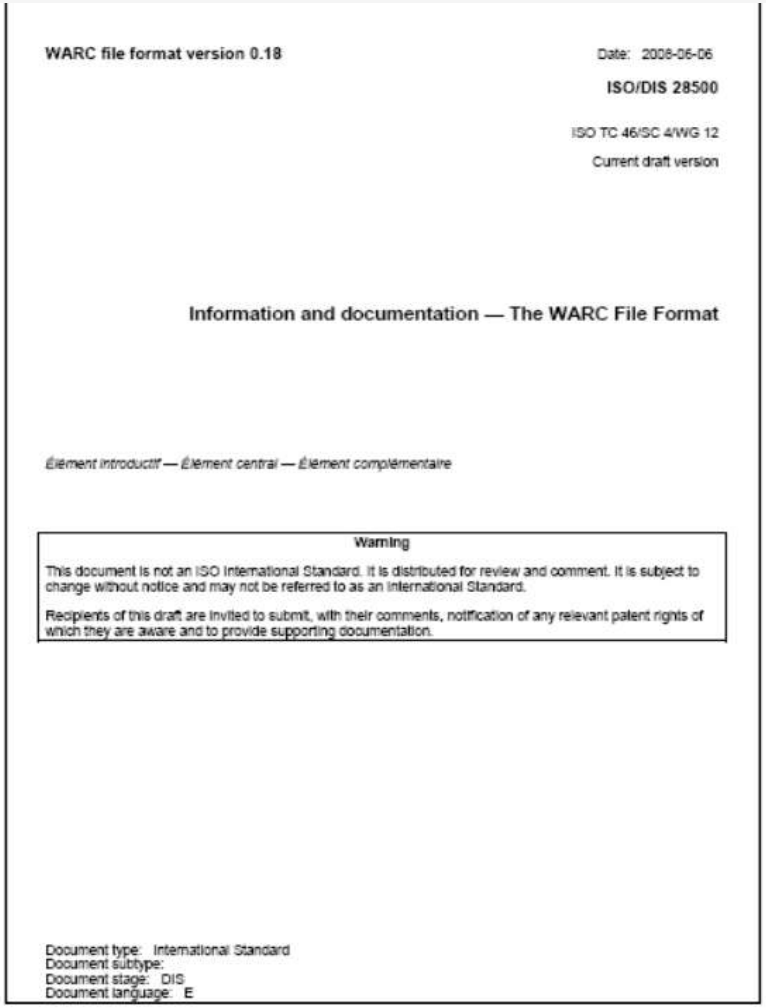
WARC文件格式是唯一面向网络资源长期保存的资源保存格式，得到了广泛的应用，具有重要的研究意义。WARC格式具有软件生态环境完善、内容丰富、便于管理、易于扩展、支持大容量文件保存等特点，同时适合网络资源和数字资源的长期保存使用。

WARC格式已经被用于大规模的网络资源的保存与交换，其内容格式的迁移和在其他类型数字资源长期保存中的应用是目前的研究热点。

网事典藏项目相关背景及建设指导思想

网络存档的标准

- WARC 格式
- WARC = Web ARChive file format
- 网络信息资源的保存格式
- 大文件格式，内嵌元数据的对象格式
- ISO 28500：2009



WARC file format version 0.18

Date: 2008-06-06

ISO/DIS 28500

ISO TC 46/SC 4/WG 12

Current draft version

Information and documentation — The WARC File Format

Élément introductif — Élément central — Élément complémentaire

Warning

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Document type: International Standard
Document subtype:
Document stage: DIS
Document language: E

网事典藏项目相关背景及建设指导思想

资源采集规则及流程

网事典藏项目建设以扬州市行政区域内的政府网站的采集和存档为重点，主要采集反映扬州市的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），提交给省馆国图进行审核确认。

网事典藏元数据加工著录规则

严格按照《推广工程数字资源联合建设政府网站元数据著录规则（2015）》对采集到的政府网站进行元数据制作，每个采集结果对应一条完整的元数据。在唯一标识符系统中注册CDOI。网事典藏机构著录项主要有：网站地址、网站名称、网站其他名称、机构名称、机构级别、机构编码、关联、网站介绍等字段。

资源采集及数据存储格式

每个网站单独全面采集。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。采集的内容全部保存为WARC1.0标准和格式。确保不含病毒、垃圾文件及采集列表外的其他信息。将采集到的文档（WARC文档）数据进行索引后发布，实现页面内容正常打开，与原网站保持一致。

数据提交、验收、保存和维护

对象数据保存在我馆提供的专用服务器上，将元数据EXCEL表格和第三方质检报告提交给国图验收。

网事典藏项目相关背景及建设指导思想

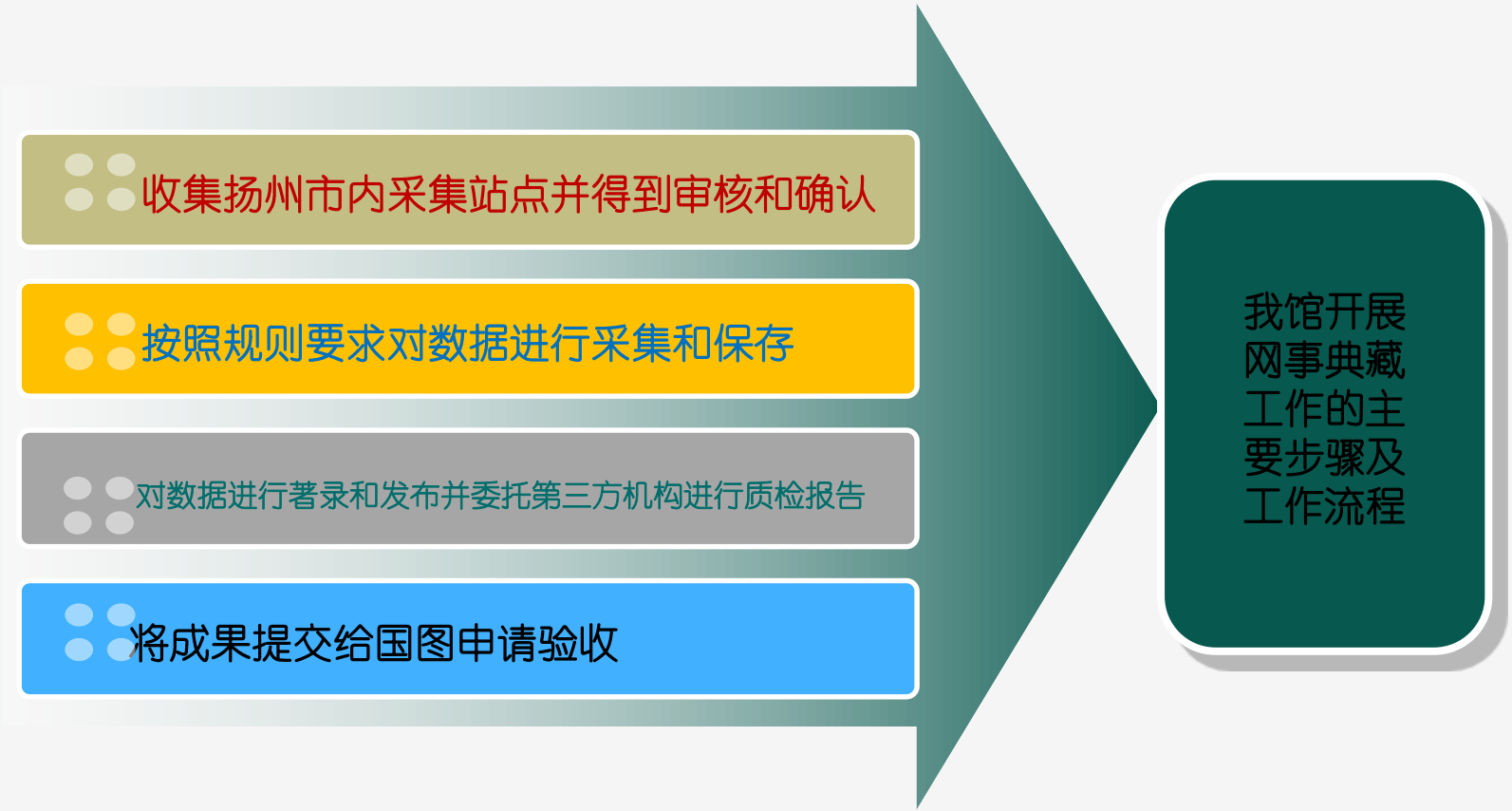
按质、按量、按时落实和完成国家数字图书馆推广工程文件的精神和项目要求。

加强项目相关的硬件投入和人力配备。

积极拓展和探索网事典藏项目在本馆的发展和应用。

02 网事典藏项目的建设流程

网事典藏项目的建设流程



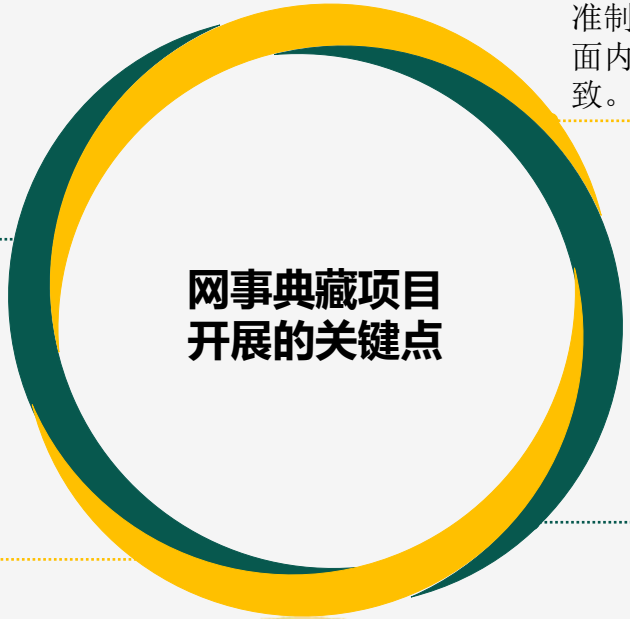
网事典藏项目的建设流程

1、部署信息采集系统

通过我馆现有服务器，部署成熟的信息采集系统，能够实现对所需站点的信息采集，包括但不限于网页内容、网页附件、网页图片等页面信息。

2、开展元数据著录

按照《推广工程数字资源联合建设政府网站元数据著录规则（2015）》对采集的网站数据资源开展元数据著录。



3、WARC制作及发布

将采集到的网站资源按照WARC1.0的相关标准制作成为WARC格式并进行保存，保证页面内容都能正常打开，且与原网站保持一致。

4、站点维护及长期保存

已收集的我馆站点进行定期维护，定期制作WARC数据，同时对已采集的网站资源进行异地备份存储，建立安全机制，保障信息安全。

网事典藏项目的建设流程

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：

政府网站采集列表

序号	网站名称	网站域名	备注

- 市馆提交给省馆初审，省馆初审后，连同初审意见一同提给交国家图书馆审核，由国家图书馆出具审核意见。

2 资源采集

根据采集列表，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WARC1.0标准，不含病毒、垃圾文件及采集列表外的其他信息。每个网站单独采集。

采集结果：WARC文档

3 数据发布

- 将采集到的文档（WARC文档）数据进行索引后收布。
- 保证页面内容都能正常打开，与不原网站保持一致。
- 数据需要在推广工程专用网络内收布。

网事典藏项目的建设流程

3 数据发布

网事典藏项目的建设流程

3 数据发布

4 元数据制作

- 《推广工程数字资源联合建设政府网站元数据著录规则》
- 每个采集结果对应一条完整的元数据。
- 需要在唯一标识符系统中注册**CDOI**。
- 将元数据制作成**excel**表。（作为成果提交）

网事典藏项目的建设流程

推广工程数字资源联合建设政府网站元数据著录规则（2015）

术语	必备性	著录内容
加工编号	必备	著录元数据的一个明确标识，定长为15位。具体组成是：资源类型代码（1位，网站：W）、采集机构代码（4位，数值取自机构代码，2-5字符位）、采集年（4位，6-9字符位）、流水号（6位，10-15字符位）。流水号应顺序排列，不同存档资源流水号不可重复。 例如：W01002015000001
CDOI	必备	著录所采集网站的唯一标识号。
网站名称	必备	著录网站名称。信息源取自网站页面首页源代码中的<title>。若<title>为空，或不反映网站内容，可用网站其他位置明显反映网站内容的名称。
网站其他名称	必备	统一著录“××网站”，对网站名称进行解释说明。例如：朝阳区人民政府网站著录的网站其他名称为“北京市朝阳区人民政府网站”。
摘要	必备	著录网站内容的总结概括性文字。摘要字数要求200字以内。
关键词	必备	著录体现网站主要内容的名词或名词短语。如有多个关键词，以半角分号间隔。
资源类型	必备	著录所保存资源的类型。统一著录为“网站”。

网事典藏项目的建设流程

术语	必备性	著录内容
内容形式	必备	著录内容形式及内容限定。参考国家标准GB/T 3469—2013《信息资源的内容形式和媒体类型标识》取值。
媒体类型	必备	著录用以承载资源内容的载体类别。参考国家标准GB/T 3469—2013《信息资源的内容形式和媒体类型标识》取值。网络信息保存资源媒体类型统一著录为“电子”。
语种	必备	著录网站的3位语种代码，可参考《新版中国机读目录格式使用手册》。如有多个语种，以半角分号间隔。
保存格式	必备	著录所采集的网站资源存档格式。统一著录为“WARC”。
机构名称	必备	著录网站的所属机构名称。著录时应以通用性、惯用性为选取原则如网站中出现多个不同的名称，选择网站最显著位置的名称。
行政级别	有则必备	著录机构所属行政级别，取值：“中央”、“省（副省）级”、“市（地）级”、“县（区）级及以下”。

网事典藏项目的建设流程

术语	必备性	著录内容
关联	有则必备	著录与当前资源存在某种关系的其他资源。
访问方式	必备	著录资源可以提供服务的范围，取值：互联网访问、推广工程专用网络访问等。
采集日期	必备	著录网站采集的日期。如果在审核过程中需重新采集，应对本项内容进行修改。
发布日期	必备	著录存档资源发布的日期。
采集地址	必备	著录政府网站的原始访问地址。
发布地址	必备	著录存档资源的发布地址。
附注	有则必备	凡未在其他著录项中著录而又有必要进一步补充说明的内容，均可著录于本项。
数据提交单位	必备	著录承建馆的名称。
所属任务年份	必备	著录联建工作的任务年度，2015年度数据则著录2015。

网事典藏项目的建设流程

元数据做成Excel表格提交

CDOI	网站名称	网站其他名称	摘要	关键词	资源类型	内容形式	媒体类型	语种	保存格式	机构名
108.ndlc	扬州市仲裁委员会 - 中[	扬州市仲裁委员会网站	该网站为扬州市仲裁委员会	扬州;仲裁	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市文化遗产网 - 中[	扬州市文物局网站	该网站为扬州市文物局网站	扬州;文物	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	中国扬州门户网站群-欢	扬州市人民政府网站	该网站为扬州市人民政府网	扬州;政府	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市发展改革委员会	扬州市发展改革委员会网站	该网站为扬州市发展改革委	扬州;发展	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市物价局 - 中国扬;	扬州市物价局网站	该网站为扬州市物价局网站	扬州;物价	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市商务局 - 中国扬;	扬州市商务局网站	该网站为扬州市商务局网站	扬州;商务	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市交通运输局 - 中[	扬州市交通运输局网站	该网站为扬州市交通运输局	扬州;交通	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市质量技术监督局 -	扬州市质量技术监督局网站	该网站为扬州市质量技术监	扬州;质量	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市食品药品监督管理	扬州市食品药品监督管理局网	该网站为扬州市食品药品监	扬州;食品	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市政府信息资源管理	扬州市政府信息资源管理中心网	该网站为扬州市政府信息资	扬州;政府	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市科学技术局 - 中[	扬州市科学技术局网站	该网站为扬州市科学技术局	扬州;科技	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市财政局-中国扬州门	扬州市财政局网站	该网站为扬州市财政局网站	扬州;财政	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市审计局 - 中国扬;	扬州市审计局网站	该网站为扬州市审计局网站	扬州;审计	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市粮食局 - 中国扬;	扬州市粮食局网站	该网站为扬州市粮食局网站	扬州;粮食	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市工商行政管理局	扬州市工商行政管理局网站	该网站为扬州市工商行政管	扬州;工商	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州国税局	扬州市国家税务局网站	该网站为扬州市国家税务局	扬州;国税	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	江苏省扬州地方税务局	扬州市地方税务局网站	该网站为扬州市地方税务局	扬州;地税	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市农业委员会 - 中[	扬州市农业委员会网站	该网站为扬州市农业委员会	扬州;农委	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市水利局 - 中国扬;	扬州市水利局网站	该网站为扬州市水利局网站	扬州;水利	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市农业资源开发局 -	扬州市农业资源开发局网站	该网站为扬州市农业资源开	扬州;农业	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市司法局 - 中国扬;	扬州市司法局网站	该网站为扬州市司法局网站	扬州;司法	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市城乡建设局-中国扬	扬州市城乡建设局网站	该网站为扬州市城乡建设局	扬州;城乡	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市规划局-中国扬州门	扬州市规划局网站	该网站为扬州市规划局网站	扬州;规划	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	城管局 - 中国扬州政府门	扬州市城市管理局网站	该网站为扬州市城市管理局	扬州;城市	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市国土资源局 - 中[	扬州市国土资源局网站	该网站为扬州市国土资源局	扬州;国土	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市住房保障和房产管	扬州市住房保障和房产管理局网	该网站为扬州市住房保障和	扬州;住房	网站	多种内容	电子	chi	WARC	扬州市
108.ndlc	扬州市民防局（扬州市人	扬州市民防局网站	该网站为扬州市民防局网站	扬州;民防	网站	多种内容	电子	chi	WARC	扬州市

5 数据验收

- 元数据审校：对编目完整的元数据按照《著录规则》进行审校，保证各字段的准确、完整。
- 对象数据审校：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。
- 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）

6 数据维护和长期保存

本馆负责对本机构制作及收布的信息及其收布网站进行长期维护，保障数据准确无误，显示正常，同时做好数据备份不长期保存工作。

03 网事典藏项目的建设成果

网事典藏项目的建设成果

已经实现对扬州市级、区县级、乡镇级242个党政机构， 241家机构官方网站的采集和著录。且实现对所有网站每一个季度采集保存一次。

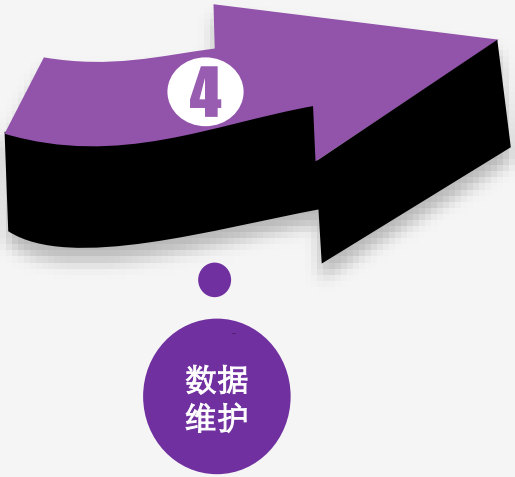
网事典藏项目的建设成果

合计采集549623个网页，资源容量达99.50G。

网事典藏项目的建设成果

按照要求已经完成了数据质检和第三方质检报告的出具。

网事典藏项目的建设成果



将对象数据保存发布在扬州图书馆服务器上，并开发了网事典藏成果发布平台web版和触摸屏版。为馆内发布和应用做积极探索和准备。

网事典藏项目的建设成果

最新典藏

扬州市中级人民法院 [1次典藏]

扬州市政协委员会 [1次典藏]

扬州市人大常委会 [1次典藏]

扬州市人民政府 [1次典藏]

宝应县妇联 [1次典藏]

中共宝应县委老干部局 [1次典藏]

宝应县总工会 [1次典藏]

中共宝应县委党校 [1次典藏]

网事典藏项目的建设成果

典藏单位 / 扬州市人民政府

历史典藏

典藏

本次典藏页面，资源容量3.02GB

加工编码	W10042015000013
CDOI	108.ndlc.35.3201009031010001/T8F45.W10002015000436
网站名称	中国扬州门户网站群-欢迎您!
网站其他名称	扬州市人民政府网站
摘要	该网站为扬州市人民政府网站，由扬州市人民政府主办，主要包括魅力扬州、信息公开、政务大厅、便民服务、互动交流、智慧门户、站群导航等栏目。
关键词	扬州;政府
资源类型	网站
内容形式	多种内容形式
媒体类型	电子
语种	chi
保存格式	WARC
机构名称	扬州市人民政府
行政级别	市（地）级
关联	
访问方式	互联网访问

04 网事典藏项目的经验分享

网事典藏项目的经验分享

01

详细、准确了解项目相关的细节及数据标准要求。

与南图和国图多请教、沟通和交流。

安排专人负责此项目的落实和实施，

定期了解掌握项目实施进度和存在的问题。

02

网事典藏项目的经验分享

03

积极投入服务器等必备资源对数据成果进行科学保存和发布应用。

充分认识网事典藏项目对扬州市政府网站资源长期保存和应用的价值和意义，积极探索新的服务和应用模式。

04

05 项目建设中遇到哪些主要问题

项目建设中遇到哪些主要问题

1.能够采集成功的网站数量少于申报数量，这就可能需要新增机构来补足。市级馆可以增收下属区县和乡镇相关机构网站。如果增加后数量还是不够，只能向国图申请一个网站间隔一个季度后采集著录两次的方案。

2.采集过程中会遇到js加载的问题导致图片或文字不显示，或现有采集工具无法解决的问题（这些问题可待国图反馈后再沟通修改）。

网事典藏项目建设中遇到哪些主要问题

3. 数据最好在图书馆保存两份，以实现灾备保存的效果。服务器的准备要考虑到今后增量采集和多次采集的配置需要。

感谢各位的聆听，谢谢！