网络资源采集工作流程及元数据规范

国家图书馆数字资源部
李一秀
2018.12

目录

一、概述

1 概念

- 网络信息采集（Web Archive，简称“WA”）
- 网络资源采集与存档：是通过使用网络爬虫软件在特定的时间按照约定的标准对网络资源进行抓取，将抓取结果进行长期保存，并将其作为镜像重新发布，为用户提供服务，使用户能够访问到网络资源在采集时刻的原貌。

一、概述

1 概念

- 原生性数字资源（Born-digital）：产生时即以数字形态出现。包括网站、论坛等一切在数字环境下诞生的资源。
- 网络资源：原生性数字资源一种典型形式。网络资源的类型包括文本、音频、视频、图像等等多种形式。

一、概述

2 重要意义

网络信息资源在现代生活中扮演越来越重要的角色，越来越多的信息以WEB形式发布。与传统的信息相比，网络资源具有数量多、更新迅速以及易逝性，特点，每天都有海量有价值的信息在消亡。

- 传承人类文化遗产
- 历史价值
- 学术意义

一、概述

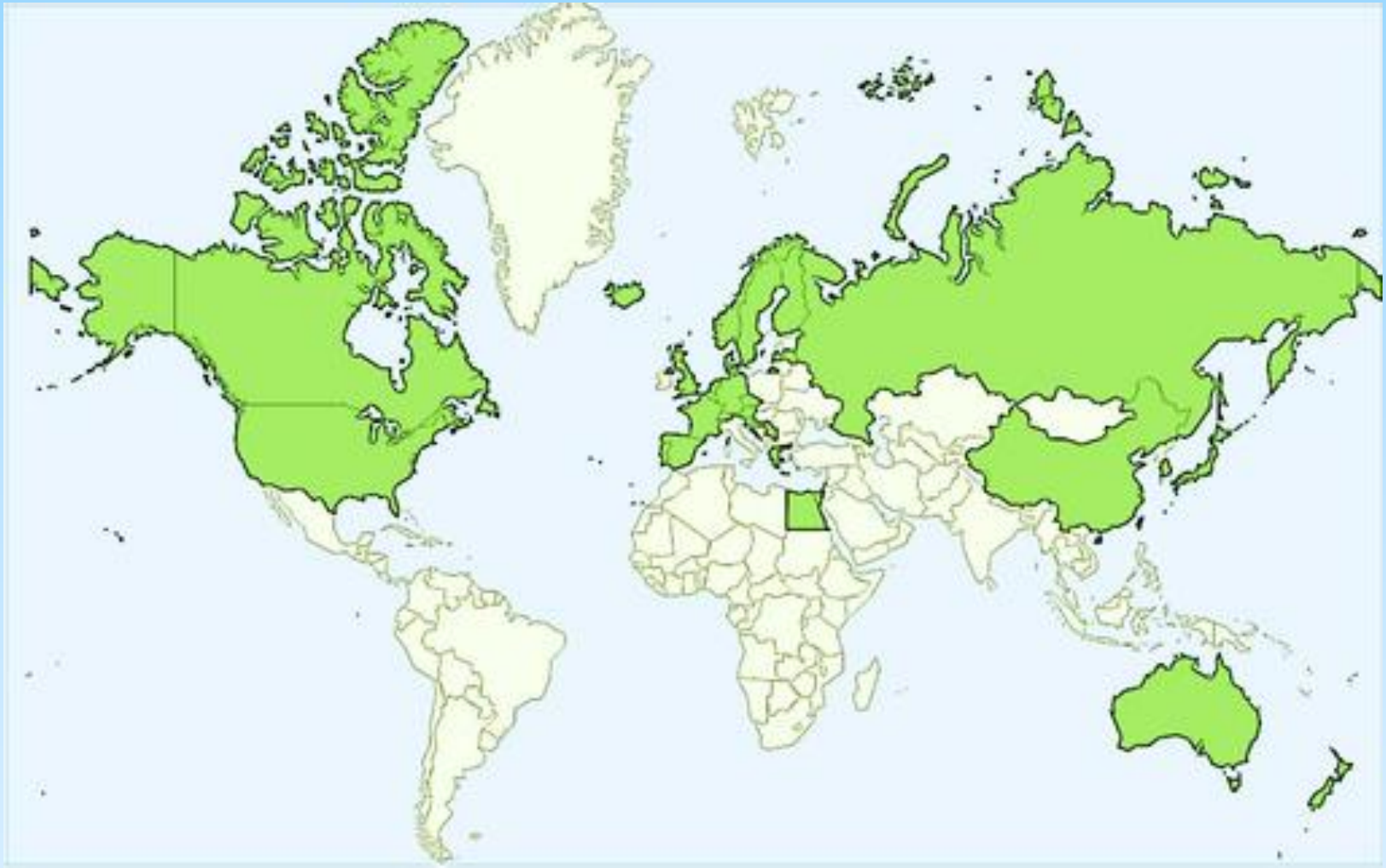
2 背景介绍

1996年，Internet Archive(IA)成立，标志网络信息资源保存研究的开始。

2003年，国际互联网保存联盟（IIPC）成立，以全面收集和保存全球互联网内容，促进通用工具、技术、标准研发和应用，解决互联网信息采集和保存为目标。

2007年，中国加入IIPC，正式成为成员之一。

二、国外主要项目介绍



二、国外主要项目介绍

国家	WA项目	牵头机构	启动时间
美国	Internet Archive（IA）	美国互联网档案馆	1996
	Minerva	美国国会图书馆	2000
法国	BnF Internet Archives	法国国家图书馆	2002
英国	UK Web Archive	英国国家图书馆	2004
	UK Government Web Archive	英国国家档案馆	1997

二、国外主要项目介绍

国家	WA项目	牵头机构	启动时间
澳大利亚	PANDORA	澳大利亚国家图书馆	1996
新西兰	New Zealand Web Archive	新西兰国家图书馆	1999
日本	WARP	日本国立国会图书馆	2004
韩国	OASIS	韩国国家图书馆	2000
加拿大	Government of Canada Web Archive	加拿大国家图书馆与档案馆	2005
挪威	Web Archive Norway	挪威国家档案馆	2001

三、国内主要进展

WA项目	牵头机构	启动时间
Web Informall	北京大学	2002
WICP	中国国家图书馆	2003
Web Archive Taiwan	台湾图书馆	1996
台大图书馆网站典藏库	台湾大学图书馆	2006

三、国内主要进展

北京大学Web Infomall项目

Web Infomall项目（中国网页信息博物馆）是北京大学在国家973和985项目支持下，由北京大学计算机系网络与分布式系统实验室建设的一个研究性项目。

专为搜集、组织与服务海量网页而设计。从2001年到2011年，该系统收集了约85亿网页，平均每天采集100-200万个网页。

（1）浏览历史网页

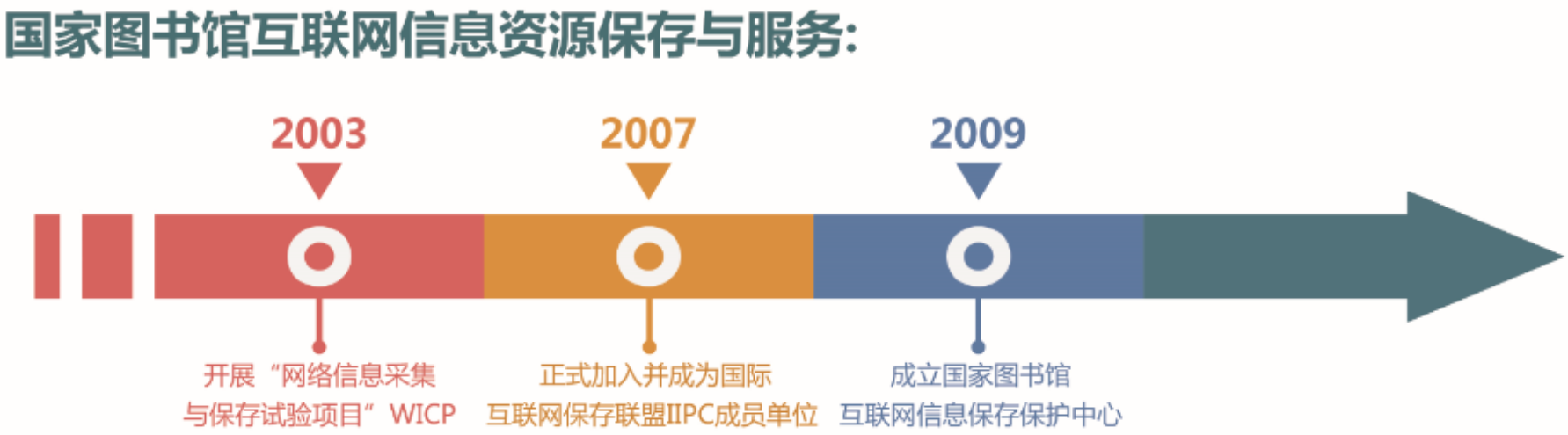
（2）历史事件专题回放

（3）数据分享

三、国内主要进展

国家图书馆**WICP**项目

- WICP (Web Information Collection and Preservation)
- 目的是对中文互联网资源进行保存保护，使反映中华文化与文明的重要互联网资源不会随着时间的流逝而消失。



三、国内主要进展


中国国家图书馆 · 中国国家数字图书馆
国家图书馆互联网信息保存保护中心
首页
关于中心
WICP
存档资源服务
专题存档
知识库
代存档服务
输入关键词,为空显示所
中心简介
互联网资源的价值已经为世界各国所公认，由于其易逝性的特点，对其进行保护已迫在眉睫。
国家图书馆互联网信息资源保存保护中心是中国国家图书馆成立的承担中国互联网信息资源长期保存保护职能的机构。
国家图书馆于2003年初成立网络文献收集与保存试验小组，通过网络信息资源采集与保存试验项目（Web Information Collection and Preservation，WICP），对互联网资源的采集与保存进行相关实验研究。随着业务的发展，2009年国家图书馆互联网信息资源保存保护中心成立。中心以全面保存中文互联网资源为目标，致力于推动中文互联网资源保存保护技术的发展与合作体系的建立，希望通过广泛的合作，实现网络采集的共建共享，促进中国互联网信息资源长期保存工作高效有序发展。
中心职责
1. 对中文网络资源进行持续保存与服务。
2. 持续跟踪与研究网络信息资源采集与保存的技术和方法，不断改进中文网络资源采集与保存的技术与环境。
3. 联合国内的公共图书馆、档案馆等存档机构，推动中文网络资源采集与保存业务在国内的发展，尽可能的完整保存中文网络信息资源。
4. 发展基于网络存档的多种应用，为中华民族的数字文化遗产的保存保护提供经验借
公告
更多
网页资源获取系统平台开发完成
网络资源采集与数字资源长期保存学
新闻 News
更多
在线选举运动的保存
新的网络存档：中国当前时事：流行
e-Helvetica的BETA版本可供...
专题 推荐
更多
两岸三通
资源 统计
更多
政府网站（采集大小）
259
908
0
7700
3600
1413
2005年
2006年
2007年
2008年
2009年
2010年
1 2 3 4 5 6

三、国内主要进展

1 资源采集

采集范围：对国内热点专题（重大会议、文化传承、科技、环境保护等主题）、国内主要政府网站（中央、省/直辖市、市、区/县等政府网站）、及国外重要网站（包括教育、航空、环境、政府与政治、经济、文化机构、军事、历史、地理、交通、团体、科技等领域）信息进行采集与保存，同时完成上述采集数据的整理、回放、索引和本地化服务。

三、国内主要进展

1 资源采集

建设模式：采

形成以国家图书馆为

全国市/区（县）级

国家图书馆

关于开展数字图书馆推广工程2018年度数字资源联合建设（首批）工作的通知

图书馆：

为进一步加强数字图书馆推广工程（以下简称“推广工程”）惠民服务的资源保障力度，2018年推广工程将继续开展数字资源联合建设工作。按照文化部要求，国家图书馆下发了《2018年数字图书馆推广工程建设任务（首批）立项通知》（国图函〔2017〕57号），负责数字资源联合建设工作的牵头管理与组织实施。经文化部批准，具体通知如下：

三、国内主要进展

1 资源采集

国家图书馆互联网资源保存保护中心：

负责牵头管理与组织实施，提供标准规范支持与人员培训，主导联合建设资源的终验、深度加工与整合等工作、

具体包括：拟定建设主题与方向，建立资源选择标准、采集收割机制、技术策略、数据存储格式、描述元数据、保存元数据、收割标准等方面的标准规范馆共同建设的保存与服务体系。

三、国内主要进展

1 资源采集

各省级承建馆：

负责对区域内市级承建馆数字资源联合建设工作的统筹管理与监督指导；

根据本馆技术能力和资源特色，按照推广工程资源建设相关标准规范，开展本馆数字资源建设。

三、国内主要进展

三、国内主要进展

1 资源采集

“网事典藏”项目建设内容：

1.网站采集

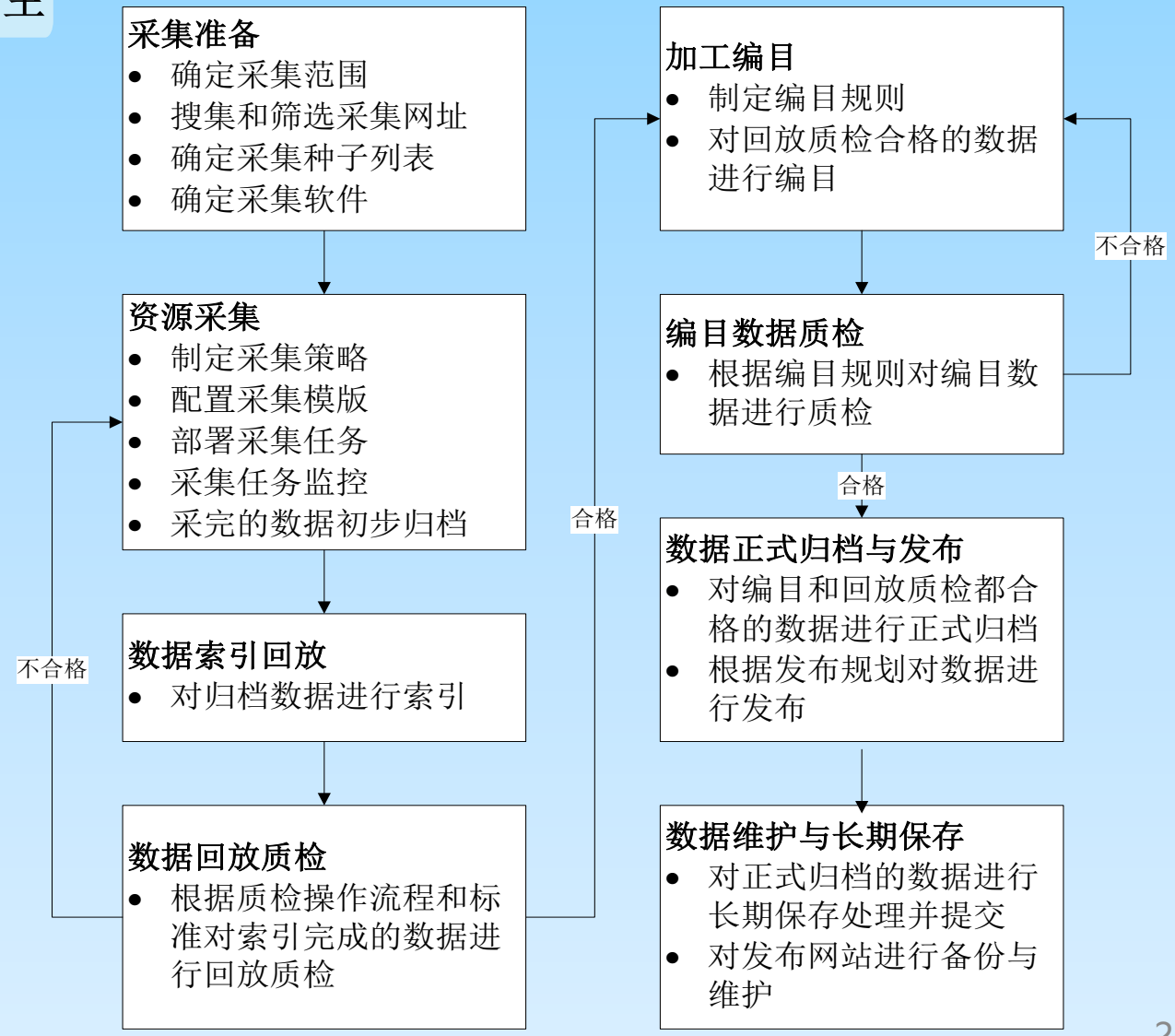
以网站的采集和存档为重点，主要采集反映所在行政区域的政治、经济、文化发展等信息的网站，整站采集。

2.专题资源采集

以专题的采集和存档为重点，主要采集反映所在行政区域当年政治、社会、文化、科技等领域的热点专题，例如：省级地方两会、洽谈会、民族节日等，全国性热点问题不予采集，如“一带一路”等。

三、国内主要进展

2 工作流程



图：网络采集保存与服务工作流程

三、国内主要进展

3 服务方式

四、网络资源采集与存档流程

一
- 采集准备

二
- 资源采集

三
- 加工编目

四
- 数据发布

五
- 数据验收

六
- 数据维护和长期保存

四、网络资源采集与存档流程

1 采集准备

- 按照网事典藏项目建设内容，对符合收录要求的网站、专题资源进行全面整理，确定采集范围。
- 网站采集：将拟采集的网站网址（URL地址）整理成采集清单（EXCEL表格）

序号	网站名称	采集地址
1		
2		
3		
…		

- 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。

四、网络资源采集与存档流程

1 采集准备

- 专题资源采集：将拟采集的专题整理成专题信息表（EXCEL表格），并将每个专题需要采集的资源整理成采集清单（EXCEL表格），由申报馆提交给国家图书馆审核，由国家图书馆出具审核意见。

推广工程数字资源联合建设网事典藏专题采集项目-采集清单（2018）

专题名称：

序号	专题编号	资源名称	资源类型	采集地址
1				
2				
3				
…				

注：专题编号、资源名称、资源类型、采集地址的著录规则和内容详见《推广工程数字资源联合建设网事典藏专题资源元数据著录规则（2018）》

四、网络资源采集与存档流程

2 资源采集

- 根据采集清单，利用**网络采集软件**，对网站进行全面采集。
- 要求所采集的文件包含采集列表中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。
- 专题资源采集要求所采集的文件包含采集清单中专题资源（网页、频道、网站）的全部内容，但不包括论坛、博客、微博、个人网页等自媒体内容以及需链接后台数据库的内容。专题中的每条资源需单独采集。
- 所采集的文档格式遵循**WARC**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。

四、网络资源采集与存档流程

3 元数据制作

- 按照元数据著录规则对采集到的网站、专题资源在指定的系统里进行元数据制作。
- 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**。
- 提交格式：将元数据制作成**Excel表**

四、网络资源采集与存档流程

4 数据发布

- 将采集到的文档（WARC文档）数据进行索引，质检合格后予以发布，要求保证页面内容都能正常打开，且与原网站保持一致。
- 采集的网站、专题须在推广工程专用网络内发布，为用户提供服务；若没有连通推广工程专用网络，须在局域网内发布。

四、网络资源采集与存档流程

5 数据验收

- 按照联建方案规定的项目进度，各馆在规定日期前，向国家图书馆提交已由第三方机构初检合格的全部数字资源。经国家图书馆终验合格后， 出具结项证明，提交成品数据。

四、网络资源采集与存档流程

5 数据验收

- 元数据审校：对编目完整的元数据按照元数据著录规则进行审校，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）

四、网络资源采集与存档流程

6 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。

四、网络资源采集与存档流程

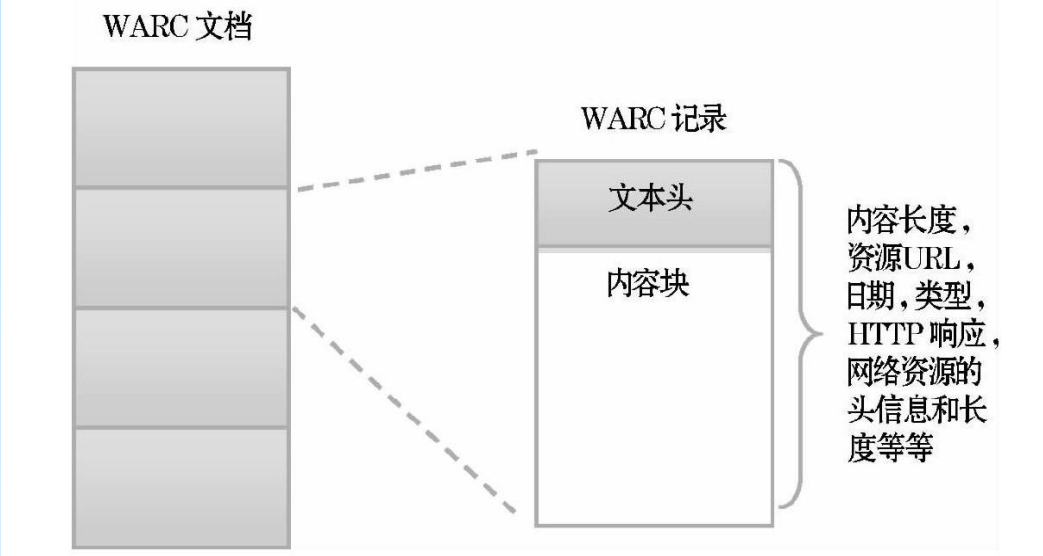
7 提交格式

- 1、元数据：元数据以excel表格方式提交。
- 2、对象数据：采集的政府网站需要在推广工程专用网络（虚拟网或专网）内发布，为用户提供服务。
- 3、第三方质检报告。

五、相关标准

1 网络存档的标准

- WARC = Web Archive file format
- WARC 格式: 网络资源存档（Web Archiving，WA）中使用的文件格式
- 大文件格式，内嵌元数据的对象格式

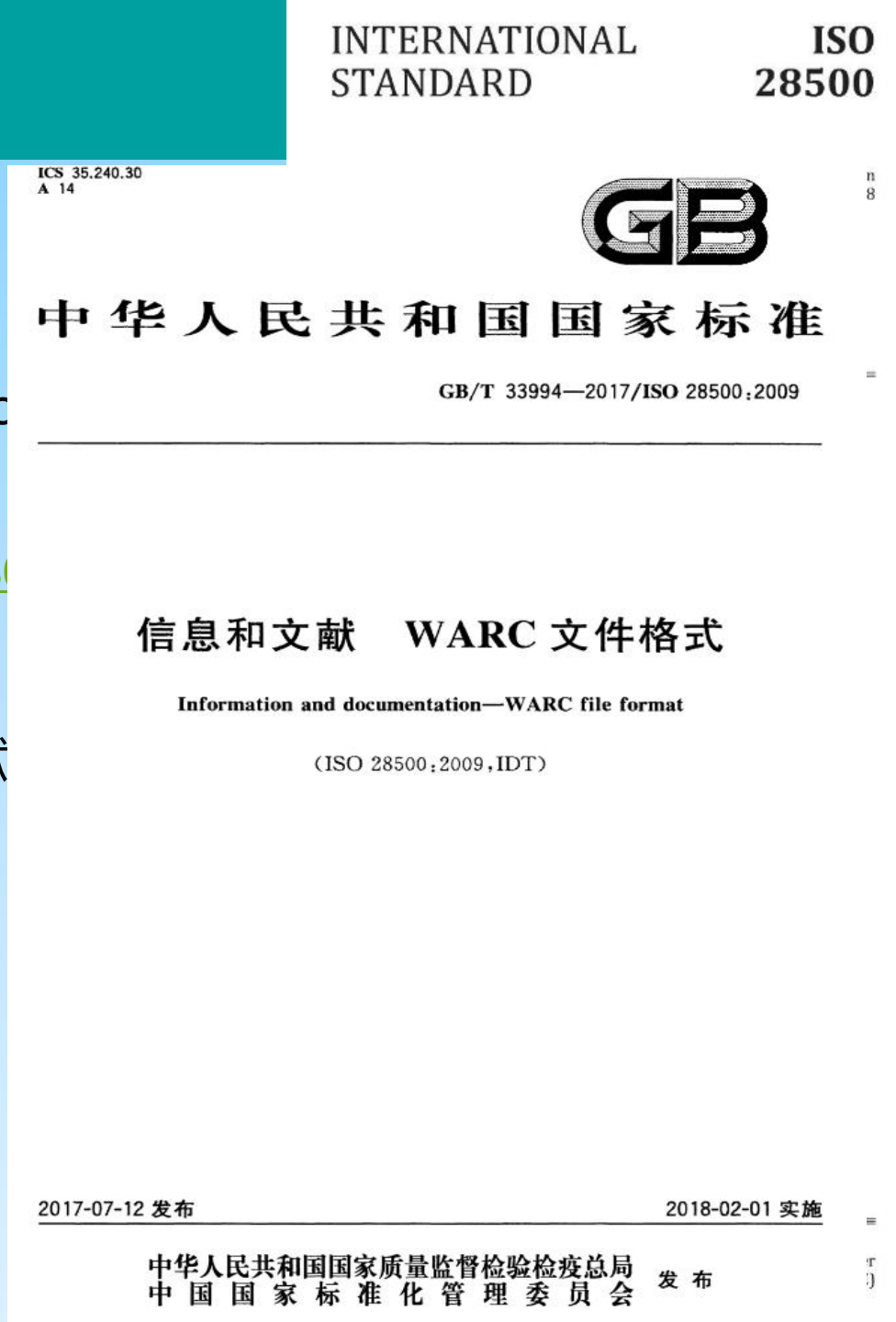


WARC 结构示意图

五、相关标准

1 网络存档的标准

- ISO 28500：2017 Informatic
  file format
  （https://www.iso.org/standard/68
- GB/T33994-2017 信息和文献



INTERNATIONAL STANDARD

ISO 28500

ICS 35.240.30
A 14

GB

中华人民共和国国家标准

GB/T 33994—2017/ISO 28500:2009

信息和文献 WARC 文件格式

Information and documentation—WARC file format

(ISO 28500:2009,IDT)

2017-07-12 发布 2018-02-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会 发布

五、相关标准

WARC文件格式的特点：

1.具备完善的软件生态环境，易于使用。

2.记录了大量的信息，保留了当时的网络环境

3.支持打包和压缩，便于管理和保存

4.支持大容量资源的保存

5.易于扩展

五、相关标准

1 网络存档的标准

WARC 文件格式是 IIPC 牵头开发的网络存档格式，在 IIPC 内部，WARC 格式中已经得到了大量应用，各成员机构都以 WARC 文档为标准保存了大量的网络资源。

UK WAC 、 BnF WA 、 Pandora 项目等也都使用WARC格式存储了大数据量的网络，美国国家档案馆也发布指南将WARC格式作为文件进馆可接受格式之一。我国的政府和档案部门也选用了WARC文件格式进行网络存档。WARC 格式在网络存档资源的建设与交换中起到了重要作用。

五、相关标准

2 资源描述标准

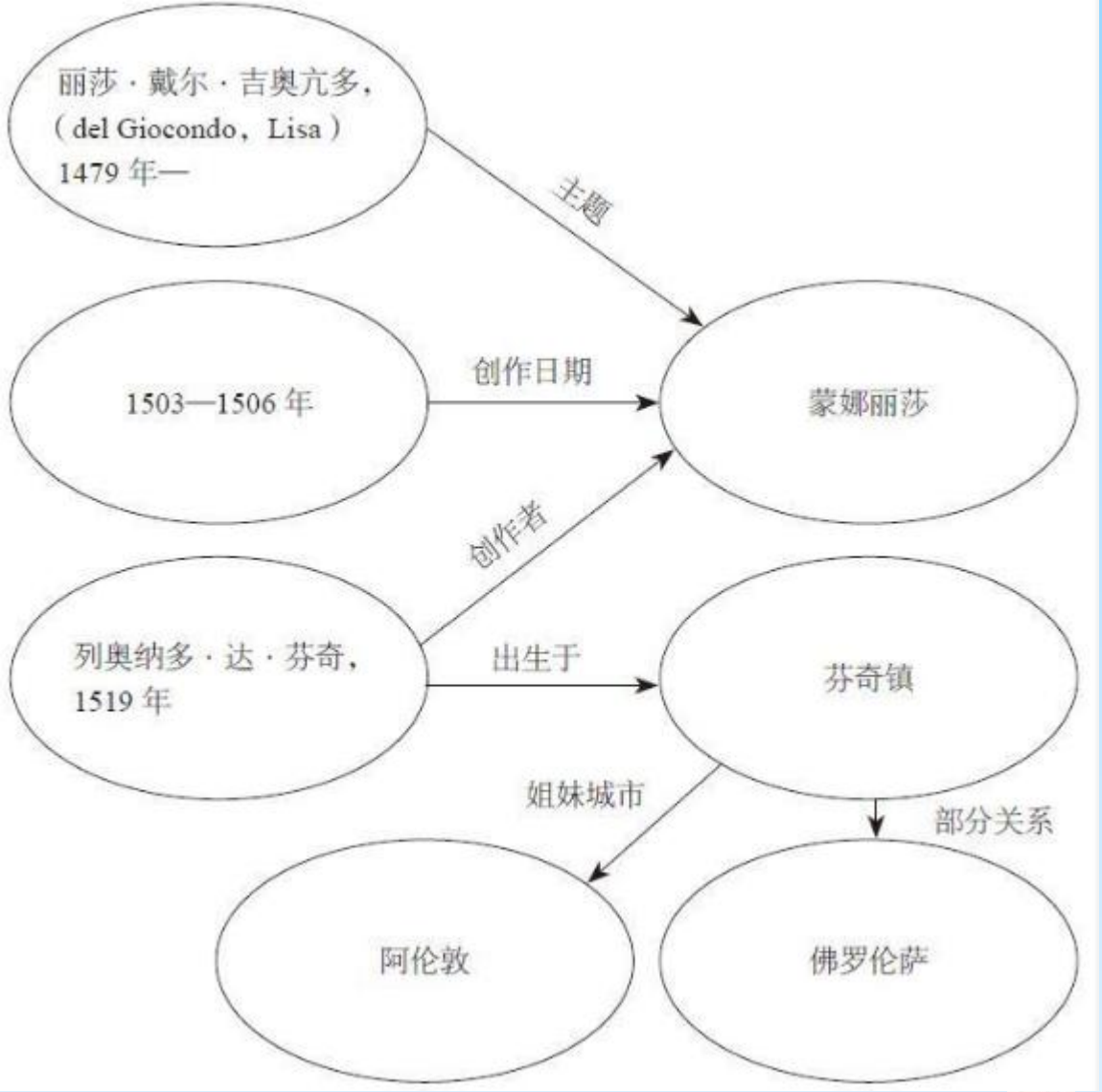
- 都柏林（DC）元数据标准
- 元数据目标著录方案（MODS）

五、相关标准

2 资源描述标准

◆元数据

一条元数据记录就是关于一个资源的主谓宾三元组的集合。



五、相关标准

2 资源描述标准

◆ 棱镜

2013年5月，美国国家安全局前外聘员工斯诺登向卫报披露国安局在本土监听机密文件。

◆ 元数据构建“大数据”

五、相关标准

2 资源描述标准

元数据对象描述模型（MODS，全称为“Metadata Object Description Schema”）：是提取MARC记录中的部分内容，用XML模式定义为一个新的元数据对象。MODS 来源于 MARC21，是用 XML 句法规则描述从 MARC21 中抽出来的元素

五、相关标准

2 资源描述标准

MODS特点：

（1）采用 XML 作为编码语言，灵活性强，对网络信息资源更加适用。

（2）采用语言标签，增加了可读性，简单实用。

（3） 适用范围广大，可作为各种资源的元数据。

（4）互操作和转换能力强。

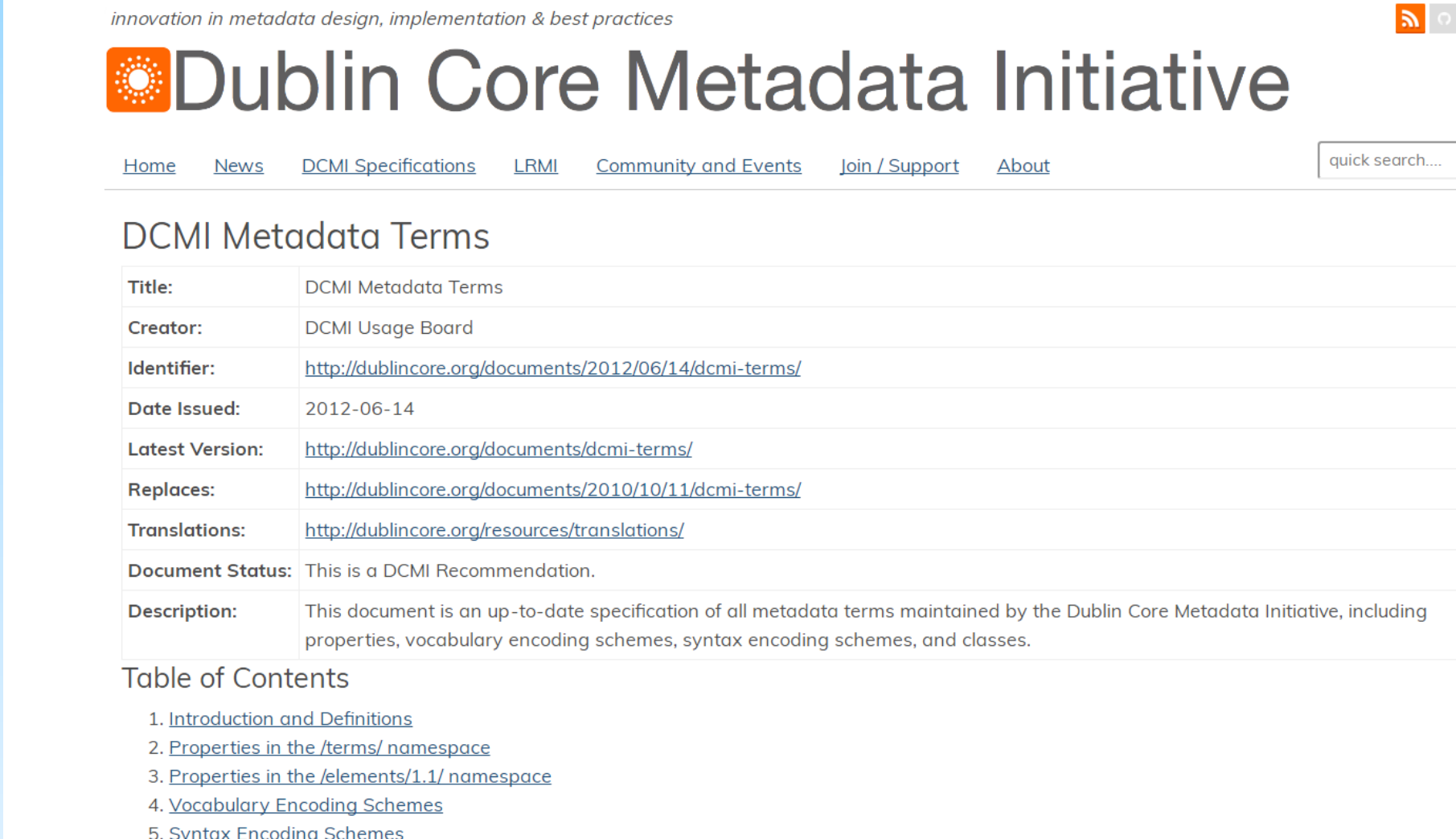
五、相关标准

LC modsxml

```
<mods xmlns="http://www.loc.gov/mods/v3" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" version="3.4" xsi:schemaLocation="http:
http://www.loc.gov/standards/mods/v3/mods-3-4.xsd">
  <identifier>lcwa00085530</identifier>
  <titleInfo>
    <title>AD2000 - a journal of religious opinion</title>
  </titleInfo>
  <typeOfResource>text</typeOfResource>
  <genre authority="marcgt">web site</genre>
  <originInfo>
    <dateCaptured encoding="iso8601" point="start" keyDate="yes">20050422</dateCaptured>
    <dateCaptured encoding="iso8601" point="end">20050502</dateCaptured>
  </originInfo>
  <language>
    <languageTerm authority="iso639-2b" type="code">eng</languageTerm>
  </language>
  <physicalDescription>
    <form authority="marcform">electronic</form>
    <internetMediaType>application/pdf</internetMediaType>
    <internetMediaType>image/gif</internetMediaType>
    <internetMediaType>image/jpeg</internetMediaType>
    <internetMediaType>text/css</internetMediaType>
    <internetMediaType>text/html</internetMediaType>
    <internetMediaType>text/plain</internetMediaType>
    <digitalOrigin>born digital</digitalOrigin>
  </physicalDescription>
  <subject authority="lcsh">
    <topic>Popes</topic>
  </subject>
  <subject authority="lcsh">
    <topic>Church history</topic>
  </subject>
  <subject>
    <name type="corporate" authority="naf">
      <namePart>Catholic Church</namePart>
    </name>
    <topic>Doctrines</topic>
  </subject>
  <relatedItem type="host">
    <titleInfo>
      <title>Papal Transition 2005 Web Archive</title>
    </titleInfo>
    <location>
      <url>http://hdl.loc.gov/loc.natlib/collnatlib.00000010</url>
    </location>
```

五、相关标准

2 资源描述标准



innovation in metadata design, implementation & best practices

Dublin Core Metadata Initiative

Home News DCMI Specifications LRMI Community and Events Join / Support About quick search....

DCMI Metadata Terms

Title:	DCMI Metadata Terms
Creator:	DCMI Usage Board
Identifier:	http://dublincore.org/documents/2012/06/14/dcmi-terms/
Date Issued:	2012-06-14
Latest Version:	http://dublincore.org/documents/dcmi-terms/
Replaces:	http://dublincore.org/documents/2010/10/11/dcmi-terms/
Translations:	http://dublincore.org/resources/translations/
Document Status:	This is a DCMI Recommendation.
Description:	This document is an up-to-date specification of all metadata terms maintained by the Dublin Core Metadata Initiative, including properties, vocabulary encoding schemes, syntax encoding schemes, and classes.

Table of Contents

1. Introduction and Definitions
2. Properties in the /terms/ namespace
3. Properties in the /elements/1.1/ namespace
4. Vocabulary Encoding Schemes
5. Syntax Encoding Schemes

五、相关标准

2 资源描述标准

“术语”属性命名域	abstract , accessRights , accrualMethod , accrualPeriodicity , accrualPolicy , alternative , audience , available ,bibliographicCitation , conformsTo , contributor , coverage , created , creator , date , dateAccepted ,dateCopyrighted , dateSubmitted , description , educationLevel , extent , format , hasFormat , hasPart ,hasVersion , identifier , instructionalMethod , isFormatOf , isPartOf , isReferencedBy , isReplacedBy ,isRequiredBy , issued , isVersionOf , language , license , mediator , medium , modified , provenance , publisher ,references , relation , replaces , requires , rights , rightsHolder , source , spatial , subject , tableOfContents ,temporal , title , type , valid
“元素集1.1”属性命名域	contributor , coverage , creator , date , description , format , identifier , language , publisher , relation , rights ,source , subject , title , type

五、相关标准

2 资源描述标准

名称：赋予数据元素的唯一标记。

URI：用于唯一标识该术语的统一资源标识符。

标签：分配给术语的标签（人类可读）。

示例

URI: http://purl.org/dc/elements/1.1/title

标签：题名(Title)

名称：title

五、相关标准

2 资源描述标准

词表编码体系	DCMIType , DDC , IMT , LCC , LCSH , MESH , NLM , TGN , UDC
语法编码体系	Box , ISO3166 , ISO639-2 , ISO639-3 , Period , Point , RFC1766 , RFC3066 , RFC4646 , RFC5646 , URI , W3CDTF
DCMI 类型词表	Collection , Dataset , Event , Image , InteractiveResource , MovingImage , PhysicalObject , Service , Software ,Sound , StillImage , Text

五、相关标准

2 资源描述标准

具体的时间

Year: YYYY (eg 1997)
Year and month: YYYY-MM (eg 1997-07)
Complete date: YYYY-MM-DD (eg 1997-07-16)

时间段：<start>/<end>

2007-03-01T13:00:00Z/2008-05-11T15:30:00Z"

GB/T 7408-2005　数据和交换格式
ISO 8601 Date and time format

五、相关标准

2 资源描述标准

◆**专题采集元数据：**著录对象为采集的网络专题资源，著录时以单次存档的专题资源为一个著录单位。

著录参考：《推广工程数字资源联合建设网事典藏专题采集项目元数据著录规则（2018）》

◆ **网站采集元数据：**著录对象为存档的网站，包括核心政府机构、事业单位、文化、艺术、科普等网站。以单次存档的网站为一个著录单位。如果一个网站具有多个主页域名，著录时作为一个对象著录。

著录参考：《推广工程数字资源联合建设网事典藏网站采集元数据著录规则（2018）》

五、相关标准

专题资源元数据

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，具体规则见《专题编号及专题资源采集加工编号命名规则》。
CDOI	有则必备	著录所采集专题资源的唯一标识号。
资源名称	必备	著录该资源的名称，一般指网络资源正式公开的名称。
所属网站	（网站频道或网页）必备	著录该资源所属网站的正式发布的规范名称。
所属专题	必备	著录该资源所属的专题编号，具体规则见《专题编号及专题资源采集加工编号命名规则》。
摘要	必备	著录该资源内容的总结概括性文字。摘要字数要求200字以内。语句简洁流畅，无语法错误。
关键词	必备	著录反映该资源主要内容的名词或名词短语。如有多个关键词，以英文半角分号间隔。

五、相关标准

术语	必备性	著录内容
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”。
采集地址	必备	著录该资源的原始访问地址。
采集日期	必备	著录该资源采集的日期。如果在审核过程中需重新采集，应对本项内容进行修改。
发布地址	必备	著录存档资源的发布地址。
发布日期	必备	著录存档资源发布的日期。
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、××图书馆局域网访问等。
中图分类	必备	著录该资源内容所属的中图分类号，多个分类号用英文半角分号间隔。
附注	有则必备	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，2018年度数据则著录2018。

五、相关标准

术语	必备性	著录内容
时间范围	有则必备	著录专题资源内容的时间特征。
空间范围	有则必备	著录专题资源内容涉及的空间特征。包括地点、地理坐标。
资源类型	必备	著录所保存资源的类型。如资源为单一网页则值为“网页”，如资源为某网站的某个专题频道，则值为“频道”，如资源为某个网站，则值为“网站”。
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	著录该资源的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，以英文半角分号间隔。

五、相关标准

内容形式	说明
图像	静态的或动态的，二维的或三维的
音乐	可以是手写的（乐谱）、演奏的、以模拟或数字形式录制的
实物	三维材料表示的内容
话语	通过人类说话声音表示的内容。
文本	通过书写词语、符号和数字表示的内容。
程序	用计算机处理或执行的数字编码指令表示的内容。
声音	通过动物、鸟类、自然噪声源，或人类声音、数字（或模拟）媒体模拟的声音而表示的内容。但不包括录制的音乐、话语录音。

五、相关标准

媒体类型词	适用的载体类型
音频	可用音频播放器播放的资源。
电子	计算机可用的资源。
缩微	可以使用缩微品阅读器的资源。
显微	使用显微镜的资源。
投影	使用投影仪的资源。
立体	可以使用立体观察器的资源。
视频	可以使用视频播放器的资源。
多媒体	用于三种或三种以上媒体类型适用的混合载体资源。
其他媒体	如果上列的词不适用于媒体类型和观看、使用或感知被著录资源内容需要的中介设备，则著录“其他媒体”这个词。

五、相关标准

网站资源元数据

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，具体规则见《网站采集加工编号命名规则》。
CDOI	有则必备	著录所采集网站的唯一标识号。
网站名称	必备	赋予资源的名称，一般指网络资源正式公开的名称。
网站其他名称	必备	统一著录为“×××网站”，对网站名称进行规范与解释说明，如果是政府网站著录，此处要求按照省、市、区县的顺序著录网站全名。例如：“朝阳区人民政府网”其网站其他名称著录应为“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。建议格式为：摘要内容分两部分，第一部分对网站内容进行整体概括，第二部分把网站包含的栏目名称列出。
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，以英文半角分号间隔。
资源类型	必备	著录所保存资源的类型。统一著录为“网站”。

五、相关标准

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，以英文半角分号间隔。
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	有则必备	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。
关联	有则必备	著录与当前资源存在某种关系的其他资源。

五、相关标准

术语	必备性	著录内容
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、××图书馆局域网访问等。
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，2018年度数据则著录2018。

五、相关标准

3 技术标准

- 采集工具： 

- 回放工具： 

五、相关标准

3 技术标准

H

当前很

- Austrian National Library, Web Archiving
- Bibliotheca Alexandrina's Internet Archive
- Bibliothèque nationale de France
- British Library
- California Digital Library's Web Archiving Service
- CiteSeerX
- Documenting Internet2
- Internet Memory Foundation
- Library and Archives Canada
- Library of Congress
- National and University Library of Iceland
- National Library of Finland
- National Library of New Zealand
- National Library of the Netherlands (Koninklijke Bibliotheek)
- Netarkivet.dk
- Smithsonian Institution Archives
- National Library of Israel

五、相关标准

3 技术标准

从功能角度看，Heritrix有丰富的爬虫规则, 可以根据需要灵活地调整规则让所采集的资源符合采集目标要求。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，可以同时保存多个版本，非常适合大规模的网络存档。

从开发角度，采用模块化的设计，支持用户在运行时选择适用的模块。在易用性方面， Heritrix使用了直观的管理界面，可以让用户快速了解采集任务进展。

五、相关标准

3 技术标准

Wayback Machine是互联网档案馆（Internet Archive, IA）所开发的WARC文档索引和回放软件。支持对 WARC文档中的URL进行索引和回放软件。它支持对WARC文档中的URL进行索引和回放，提供检索界面。Wayback是Web Archive领域中广为使用的存档资源访问系统，集索引、检索、再现等功能于一体，能够自动监测制定的目录实现WARC文档的增量索引，能够位用户提供基于URL的检索以访问Web Archive资源。但对浏览内容的支持还存在不足。

结语

- 建设模式：馆际合作，共建共享
- 采集流程：合理化，规范化
- 采集方式：混合式，按需选择
- 资源描述：标准化，规范化

Thanks
欢迎留下您的宝贵意见