



数字资源联合建设项目 浙江图书馆开展情况介绍

浙江图书馆
ZHEJIANG LIBRARY

政府信息公开项目

浙江省公共图书馆政府信息公开整合服务平台是浙江图书馆为满足公众对浙江省各级政府公开信息的有效利用而开发的一站式服务平台。本平台的整合对象是浙江省各级政府网站上所有政府信息公开信息，设计功能是提供对全省各级政府公开信息的组织、归类、整理编目、长期保存和全文搜索，为公众查询与利用浙江省各级政府公开信息提供便利。

目前仅装载少量的测试数据，部分类别的栏目还没有数据提供，后续我们将加快数据的采集和加载。

<http://61.175.198.133:8088/GovInfoOpen/portalController/portal>



网事典藏项目

浙江图书馆网事典藏资源库主要采集反映浙江地方政治、经济、文化发展等信息的政府网站和重要的事业单位网站，将采集到的网站通过组织、整理、编目后发布到互联网，供广大用户进行长期访问。目前我们已经采集了100个网站的资源，今后将继续逐步增加采集网站的数量。

<http://61.175.198.135:9090/>





网事典藏项目

浙江图书馆
ZHEJIANG LIBRARY

- ✓ 认真对待、仔细分析
- ✓ 工具选型、数据收集
- ✓ 校验编目、成果提交
- ✓ 数据留存、整合利用



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11月

本着对国图工作的高度重视、紧密配合的原则，我馆仔细分析了关于网事典藏工作的具体细节和工作流程。

对李云龙老师的《网事典藏项目建设标准及方法》进行了仔细的分析 and 的探讨，从而对网事典藏工作的项目概况、工业流程、成果提交、以及与政府信息公开的区别有了深刻的理解。明确了国图想要什么数据，以及我馆应该怎样提供数据、提供什么样的数据。



认真对待、仔细分析

- 每馆完成不少于**100个**网站的采集和长期保存
- 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

采集
要点



认真对待、仔细分析



认真对待、仔细分析

- 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

采集
成果



工具选型、数据收集

- ✓ 针对以上要点、流程以及成果提交，我们要选择一款“采集工具”来帮助我们完成此工作。
- ✓ 这个工具不但要求可以采集网站信息、支持WARC存储、编目、发布等操作，更重要的是需要这个工具能“符合国图要求、可靠”的减少工作量。
- ✓ 选择了一家与浙图长期合作的技术供应商，通过研究需求，共同开发，最终顺利完成了能达到国图要求的数据收集工作。



工具选型、数据收集

WAP特点

采集

多爬虫分布式采集、资源增量采集、断点续采

存储

国际文件保存标准 (WARC)、分布式并行存储、异构数据库、容灾、备份

展现

原版原貌展现、全文检索、资源导航、数据分析利用

控制

爬虫、存储、展现总控调度，权限、配置灵活体现



工具选型、数据收集

网事典藏效果

The screenshot displays the Wayback Machine interface for the URL <http://www.zjsme.gov.cn/newzjsme/index.asp>. The browser address bar shows the IP address 61.175.198.135:9080. The Wayback Machine logo and "4 captures" are visible. The date selected is April 24, 2015.

The website content includes a navigation menu with items like 首页, 发展指导, 创业创新, 融资担保, 协作交流, 网上办事, 交流互动, 网站地图, and 站内搜索. The main content area features a banner for "浙江中小企业网" (Zhejiang SME Network) and several news items under "媒体信息" (Media Information), such as "央行1.5万亿流动性驰援 实体经济能得几多活水" and "工商总局积极推进小微企业名录建设".

On the right side, there is a sidebar titled "中小企业工作职能机构" (SME Work Function Institutions) listing various departments like "省经信委中小企业发展规划指导处" and "省中小企业服务中心". At the bottom right, there is a section for "政策法规指南" (Policy and Regulation Guide) with the sub-heading "中小企业质量、标准、低碳".

校验编目、成果提交

标识符	网站名称		描述	类型	语种	格式			机构		域名	采集地址	日期	访问地址	
加工编号	CDOI	正题名	交替名称			文件格式	文件数量	文件大小	机构名称	机构级别	主页域名	子域名(重)	网站原址	采集日期	存档访问地址
GOV110	104	浙江省	浙江省	网站设置了办事指南	政	中	WA	992M	浙江省	司局级	www.zj	s.gov.	http://	201504	http://10.
GOV11000	108.	浙江安全	浙江省安	网站设置了综合监管	政府	中文	WARC	5 337M	浙江省安	司局级	ajj.zj.g	www.zjs	http://a	20150427	http://10.18
GOV11000	108.	浙江统计	浙江省统	网站设置了浙江省情	政府	中文	WARC	7 414M	浙江省统	司局级	www.zj.stats.gov		http://v	20150424	http://10.18
GOV11000	108.	欢迎光临	浙江省海	网站设置了组织机构	政府	中文	WARC	9 516M	浙江省海	司局级	hyj.zj.gov.cn		http://h	20150424	http://10.18
GOV11000	108.	浙江省机	浙江省机	网站设置了了解我局	政府	中文	WARC	5 376M	浙江省机	司局级	www.zjhq.com		http://v	20150424	http://10.18
GOV11000	108.	浙江省人	浙江省人	网站设置了政务公开	政府	中文	WARC	8 511M	浙江省人	司局级	fzb.zj.gov.cn		http://f	20150508	http://10.18
GOV11000	108.	浙江省物	浙江省物	网站设置了资讯中心	政府	中文	WARC	4 160M	浙江省物	司局级	www.zjpi.gov.cn		http://v	20150424	http://10.18
GOV11000	108.	浙江省文	浙江省文	网站设置了信息公开	政府	中文	WARC	19 1.63G	浙江省文	司局级	www.zjww.gov.cn		http://v	20150414	http://10.18
GOV11000	108.	首页-欢	浙江省地	网站设置了机构简介	政府	中文	WARC	4 133M	浙江省地	司局级	www.zjdz.gov.cn		http://v	20150427	http://10.18
GOV11000	108.	首页-	浙江省戒	网站设置了组织机构	政府	中文	WARC	3 41M	浙江省戒	司局级	www.zjlj.gov.cn		http://v	20150427	http://10.18
GOV11000	108.	宁波人大	宁波市人	网站设置了人大简介	政府	中文	WARC	5 237M	宁波市人	省部级	www.nbrd.gov.cn		http://v	20150427	http://10.18
GOV11000	108.	政协宁波	中国人民	网站设置了政协概况	政府	中文	WARC	3 114M	中国人民	省部级	www.nbzx.gov.cn		http://v	20150427	http://10.18
GOV11000	108.	温州政协	中国人民	网站设置了要闻快递	政府	中文	WARC	11 847M	中国人民	司局级	www.wzzx.gov.cn		http://v	20150427	http://10.18
GOV11000	108.	湖州政协	中国人民	网站设置了政协概况	政府	中文	WARC	1 77.5M	中国人民	司局级	zxb.huzhou.gov.cn		http://z	20150429	http://10.18
GOV11000	108.	绍兴人大	绍兴市人	网站设置了人大概览	政府	中文	WARC	2 53.9M	绍兴市人	司局级	sxrd.sx.gov.cn		http://s	20150429	http://10.18
GOV11000	108.	绍兴政协	中国人民	网站设置了政协概况	政府	中文	WARC	2 127M	中国人民	司局级	www.zjxzx.gov.cn		http://v	20150428	http://10.18
GOV11000	108.	政协金华	中国人民	网站设置了政协概况	政府	中文	WARC	7 554M	中国人民	司局级	www.jhzx.gov.cn		http://v	20150428	http://10.18
GOV11000	108.	衢州政协	中国人民	网站设置了政协概况	政府	中文	WARC	1 26.5M	中国人民	司局级	www.quzhouzx.gov		http://v	20150501	http://10.18
GOV11000	108.	首页-舟	舟山市人	网站设置了人大概览	政府	中文	WARC	2 110M	舟山市人	司局级	www.zjzsr.gov.cn		http://v	20150429	http://10.18
GOV11000	108.	舟山市政	中国人民	网站设置了舟山概况	政府	中文	WARC	1 84.8M	中国人民	司局级	www.zszx.gov.cn		http://v	20150429	http://10.18
GOV11000	108.	中国台州	台州市人	网站设置了人大概论	政府	中文	WARC	2 37.6M	台州市人	司局级	www.zjtzrd.gov.cn		http://v	20150430	http://10.18
GOV11000	108.	台州市政	中国人民	网站设置了政协简介	政府	中文	WARC	4 230M	中国人民	司局级	www.zjtzzx.gov.cn		http://v	20150430	http://10.18
GOV11000	108.	丽水政协	中国人民	网站设置了政协简介	政府	中文	WARC	19 1.75G	中国人民	司局级	www.lszx.gov.cn		http://v	20150427	http://10.18
GOV11000	108.	杭州民政	杭州市民	网站设置了政务公开	政府	中文	WARC	3 217M	杭州市民	司局级	www.hzmz.gov.cn		http://v	20150425	http://10.18
GOV11000	108.	杭州市规	杭州市规	网站设置了政务公开	政府	中文	WARC	2 152M	杭州市规	司局级	www.hzplanning.g		http://v	20150425	http://10.18
GOV11000	108.	杭州市卫	杭州市卫	网站设置了政务公开	政府	中文	WARC	4 314M	杭州市卫	司局级	www.hzwsjsw.gov.		http://v	20150425	http://10.18
GOV11000	108.	政务频道	杭州市质	网站设置了机构介绍	政府	中文	WARC	3 215M	杭州市质	司局级	www.hzqts.gov.cn		http://v	20150425	http://10.18
GOV11000	108.	杭州市人	杭州市人	网站设置了外办简介	政府	中文	WARC	3 143M	杭州市人	司局级	www.hzfa.gov.cn		http://v	20150429	http://10.18

数据留存、整合利用

浙江图书馆网事典藏资源库主要采集反映浙江地方政治、经济、文化发展等信息的政府网站和重要的事业单位网站，将采集到的网站通过组织、整理、编目后发布到互联网，供广大用户进行长期访问。目前我们已经采集了100个网站的资源，今后将继续逐步增加采集网站的数量。

数据留存：以政府网站和门户网站为采集和存档的重点，主要采集反映浙江省政治、经济文化发展等信息的政府网站或事业单位网站，采用的方法为专题存档，以网页为信息单元进行网络信息存档。按不同专题确定对象网站，从对象网站的首页开始，收集该网站下的所有网页，进行内容提取、自动分类和标引，并按照符合国际标准的格式将其保存到存档系统中进行长期保存，能够标引的数据直接拆分存入数据库，对存档系统和数据库都提供检索功能，检索结果都能指向或者还原出原始网站，并可选择部分页面保存网页快照。

整合利用：目前网事典藏项目后续工作正在按照计划有序的进行着，估计2016年春节过后将正式投入使用。



数据留存

前台部分功能



原版原貌



数据留存、整合利用

后台部分功能（原型）

- 系统管理
 - 模块管理
 - 角色管理
 - 用户管理
 - 字典管理
 - 系统日志管理
 - 业务日志管理
- 采集管理
 - 网站管理
 - 采集网站
 - 采集任务
 - 搜索引擎管理
 - 搜索引擎规则
 - 搜索引擎内容
- 数据管理
 - 网站数据
 - 编目数据
 - 网页数据
- 专题管理
 - 动态专题
 - 静态专题
 - 静态专题数据
- 统计管理
 - 网站数据统计
 - 搜索引擎数据统计
 - 动态专题数据统计
 - 静态专题数据统计
 - 个人工作量统计

操作人 🔍 查询 🔄 重置

📄 导出 🖨 打印

序号	操作人	采集网站数量	采集搜索引擎数量	创建静态专题数量	创建动态专题数量	编目网站数量	编辑网页数量
1	张国栋	98	87	76	9	76	98765
2	张三	87	76	65	8	65	87654
3	李四	76	65	54	7	54	76543
4	王五	65	54	43	6	43	65432
5	赵六	54	43	32	5	32	54321
6	刘七	43	32	21	4	21	43210

10 ⏪ ⏩ 第 1 共 1 页 🔄 显示1到2,共2记录

谢谢！

