大数据环境下的网络信息保存与利用

国家图书馆　　张 炜

目录

1

背景：大数据时代到来

A pp

B ig Data

C loud computing

一天网络中到底能产生多少数据呢？

人们将越来越多的意识到数据的重要性

红酒广告

初期投放

财经

财经频道貌似红酒目标客户
高端人士的聚集地

电商买家
动态肖像系统

数据分析

常访问**军事类博客**的网民
对红酒更感兴趣

改变策略

选取该网站博客频道中的军事类博客页面投放红酒广告
一个投放周期结束后，有效转化率高达**18%**，而行业的平均水平却低于**5%**

TrueCar之所以受到消费者的喜爱，因为只是做了一件事情：

收集全美汽车销售商的数据，利用**大数据分析**将**车价透明化**并发布在自己的信息平台

将价格从低到高分为4个区间：

罕见低价 超值价格 不错的价格 超过市场均价的价格

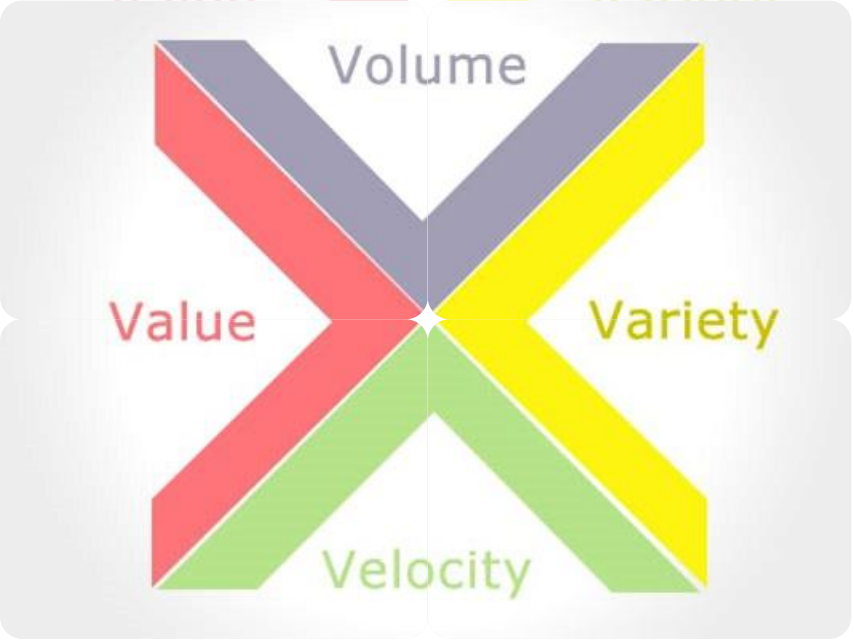
大数据的4V特征

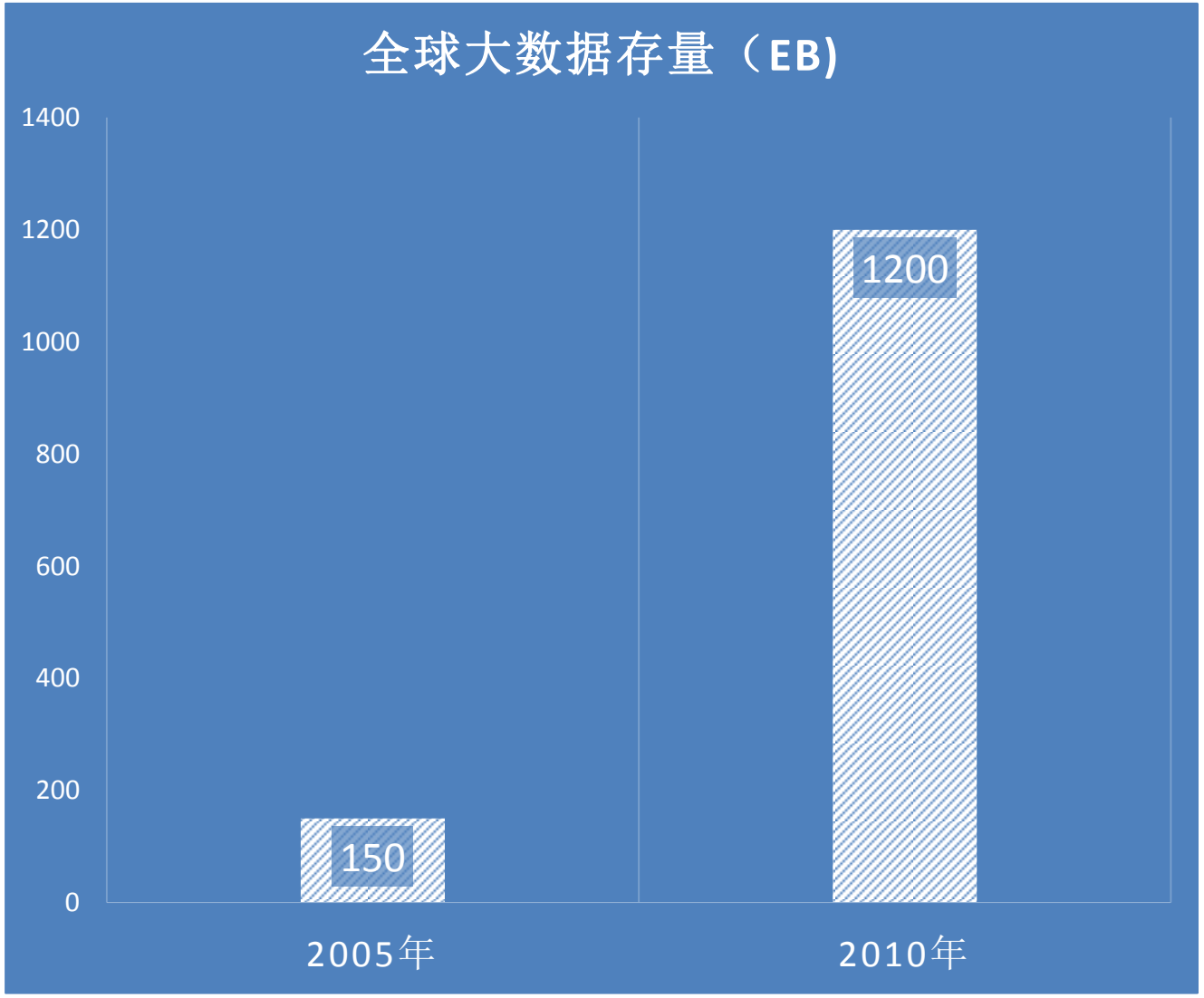
数量: 即数据巨大,从TB级别跃升到PB;

多样性:数据种类繁多,不仅包括格式化的数据,还包括来自互联网的网络日志、视频、图片、地理位置信息等；

速度：处理速度快；

真实性：追求高质量的数据；

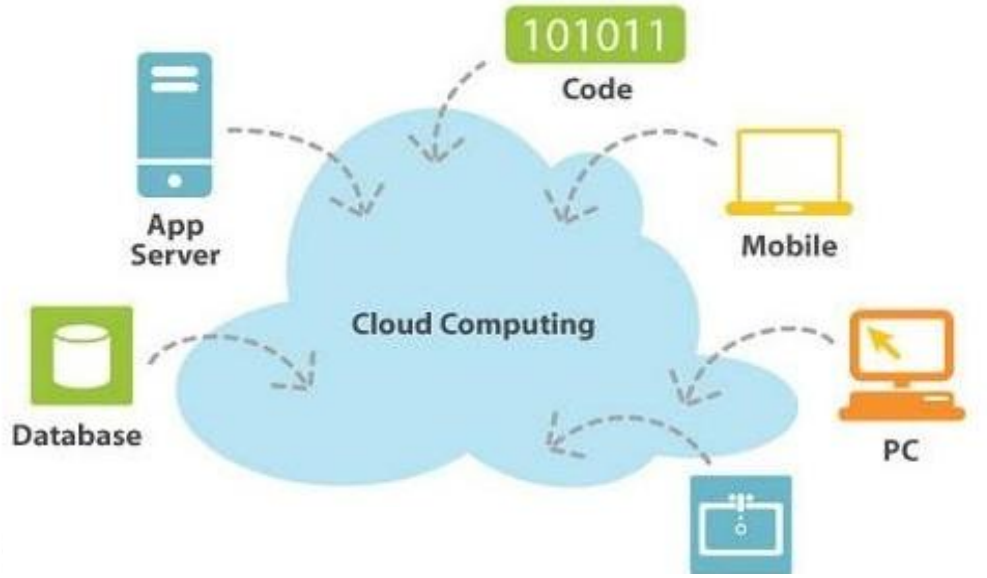




从2005年150EB增长到2010年1200EB

预计将以40%的年增长率继续增长

2020年将达到2007年的44倍

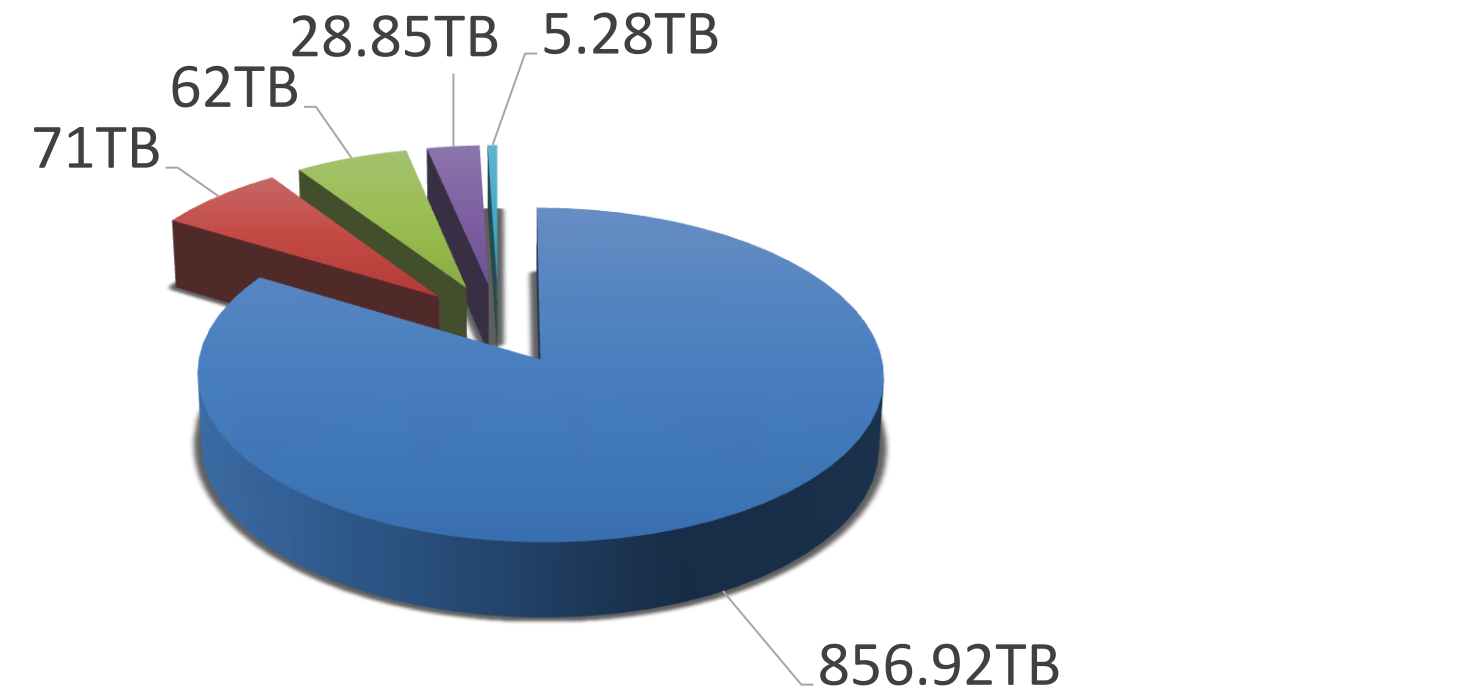


一是多种类型的**海量资源**及庞大**用户数据**；

二是图书馆的生产数据和业务数据在根据数字资源生命周期在各业务系统间**高速流转的数据体系**；

三是通过图书馆海量数据满足用户的**知识需求和个性化服务需求**的价值体现

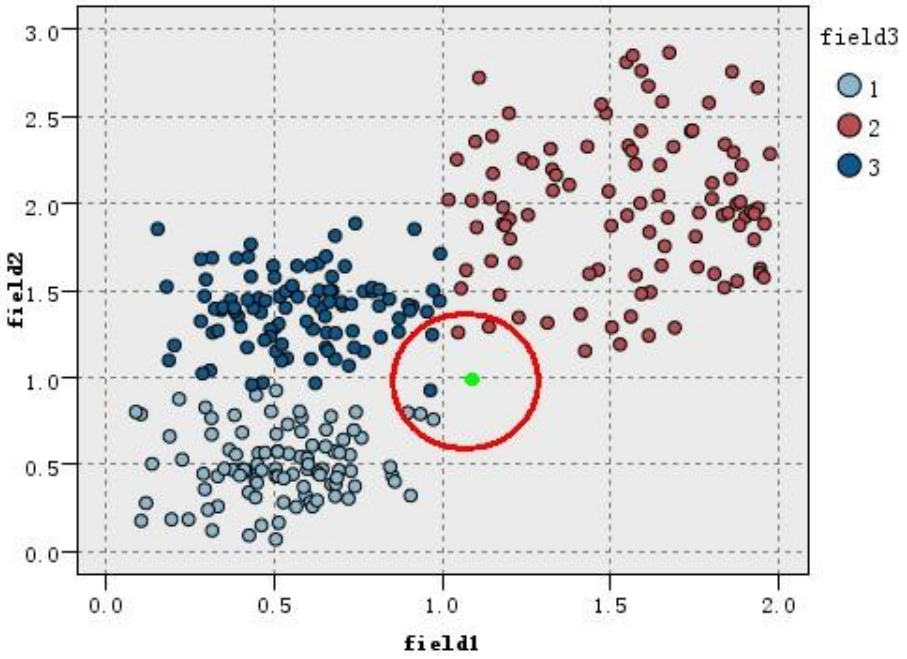
国家数字图书馆数据类型分布表	
资源数据	元数据、对象数据、书目数据、规范数据、馆藏数据、单册数据、关联数据
用户数据	用户个人信息数据、在服务中产生的不同维度的用户行为数据
业务数据	各业务系统运转的业务数据，如：采访数据、缴送数据、征集数据、咨询数据等
管理数据	图书馆运转和服务中产生的各类管理数据，如项目建设数据、合同数据、财务数据、人事数据



数字资源总量1024TB

图书馆需要使资源从数据层面的揭示与描述向数据挖掘和知识发现转变：

- 元数据的仓储式建设与统一管理
- 统一发现与服务的完善
- 利用语义和关联技术实现数字馆藏的组织和聚合
- 结合数据分析技术实现数字资源的可持续发展
- 通过扩大网络信息的保存规模，提升网络资源服务效力



2

网络信息成为重要数据资产

截至2014年12月
中国：

网站数量	335万
网民规模	6.49亿
互联网普及率	47.9%
手机网民规模	5.57亿
手机上网的网民比例	85.8%

Netcraft统计：
全球网站数量**9.92亿**个，将突破10亿大关

随着互联网高速发展，网络成为一个国家除领土、领海、领空、太空之外的**第五大疆域**

政府决策

企业经营

网络安全

人民生活

社会各领域资讯在“第五空间”的碰撞与发酵中深度融合，全面反映了国家社会各领域的真实面貌，对分析我国的发展现状并做出正确的判断和决策具有重要意义。

网络信息保存与利用在世界各国受到普遍重视

2005年 设立开放源中心
2011年 建设犹他数据中心
2012年 发布《大数据研究和发展计划》
2013年 新建另一个高性能计算中心

2014年 欧盟启动“地平线2020”计划

美国、英国、法国、澳大利亚、丹麦、日本等的国家图书馆在政府的统一管理和支持下，联合其他档案机构开展了各自的网络存档工程，并开展科学研究

网络信息保存和保护工作迫在眉睫

网络发展与更新速度惊人，网络产业的融合与重组更加剧了网络信息的更新换代，网络信息成为易逝的不可再生资源。

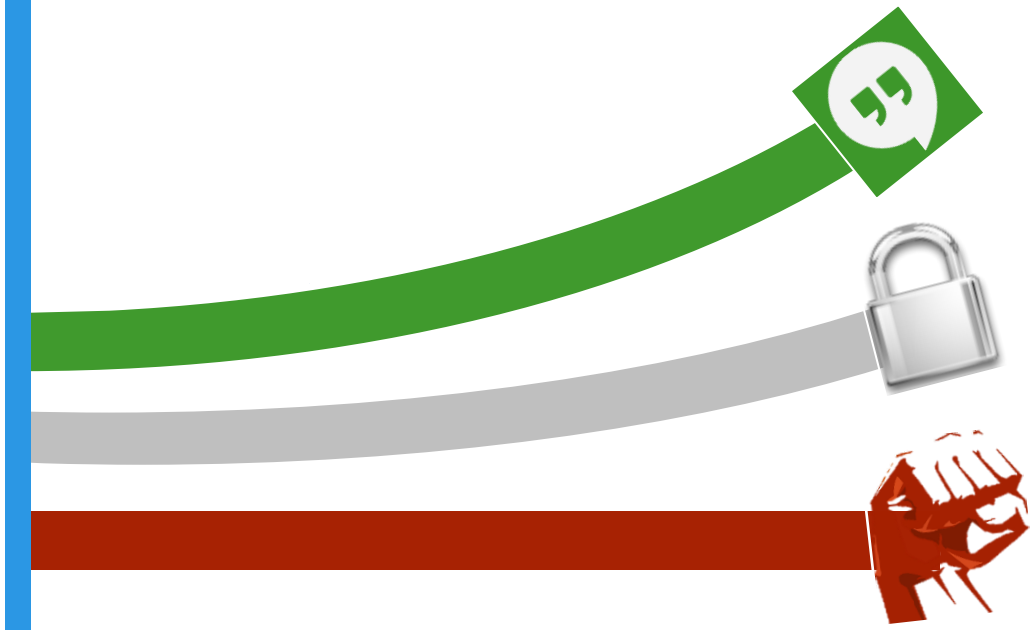
我国大量的网络资源因为得不到有效的保存和保护而流失，导致信息价值的严重浪费，对文化和文明的传承也造成了影响。

网络信息保存和保护工作迫在眉睫

加强网络治理，维护国家网络安全的需要

没有网络安全就没有国家安全。

对网络信息进行保存和研究，有利于占领网络阵地，加强网络安全治理，维护我国的网络安全、文化安全与国家安全，推动我国从网络大国向网络强国转化。



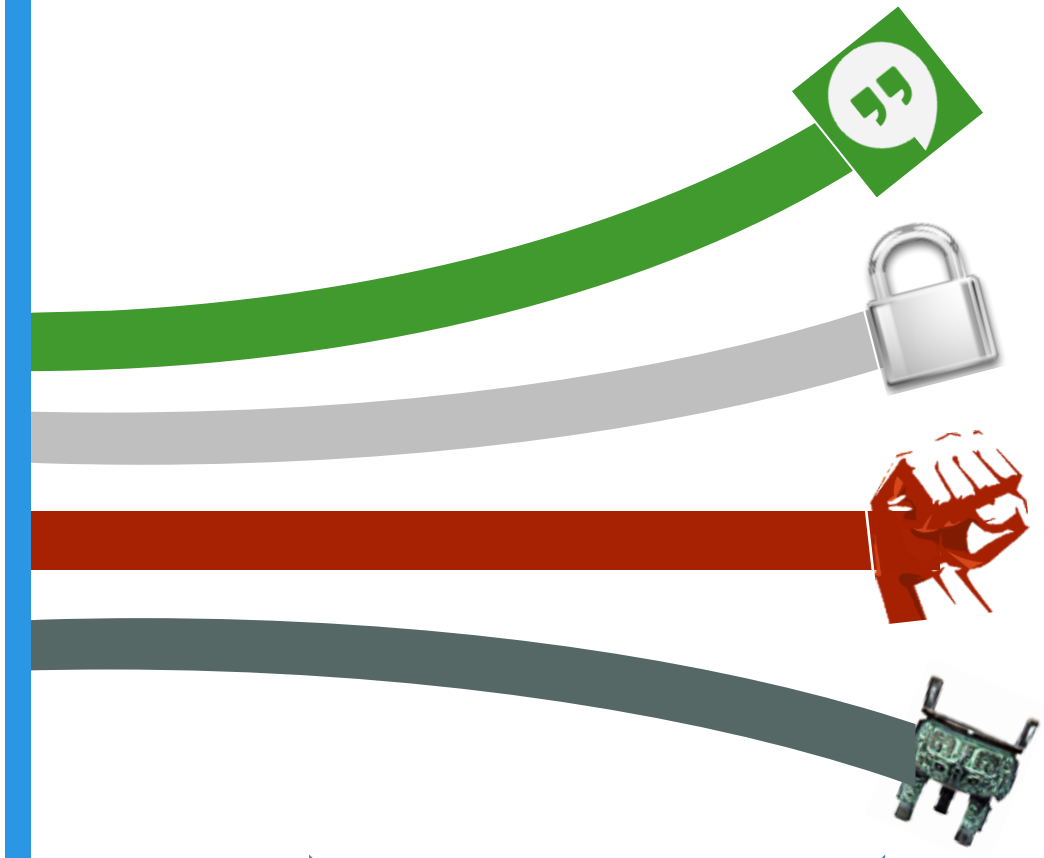
网络信息保存和保护工作迫在眉睫

加强网络治理，维护国家网络安全的需要

推进信息化建设，增强国家竞争力的需要

网络信息掌握的多少已经成为国家竞争力和创新力的重要标志。

保存网络信息，有利于促进网络基础设施建设，增强自主创新能力，发展国家信息经济，提升网络安全保障能力，不断增强国家竞争力和创新力。



网络信息保存和保护工作迫在眉睫

加强网络治理，维护国家网络安全的需要

推进信息化建设，增强国家竞争力的需要

承载中华数字记忆，提升国家文化软实力需要

有利于记录时代文明发展脉络，有利于讲好中国故事，传播中国声音，阐释中国特色，不断激发中华优秀文明的活力，有利于保障并促进中华优秀文明的广泛、久远的传播，增强中华优秀文化的辐射力与影响力


› Register / Sign in
Home
Browse
Campaigns
Content
Tags
SINGAPORE MEMORY
408,181 Memories Added
› Add Your Memory
My School Days
Our Neighbourhoods
Food Nostalgia
I Remember KTM
Singapore Story: The 3H's
‹ SPOTLIGHT
ME ›
‹ Recent Memories
irememberKKH
Science Centre
YMCA
Recent Blog Feeds ›

3

国外网络信息保存与利用
——走在前端

国外网络信息保存的发展历程

1996年Internet Archive成立标志着Web Archive实践探索的开始。

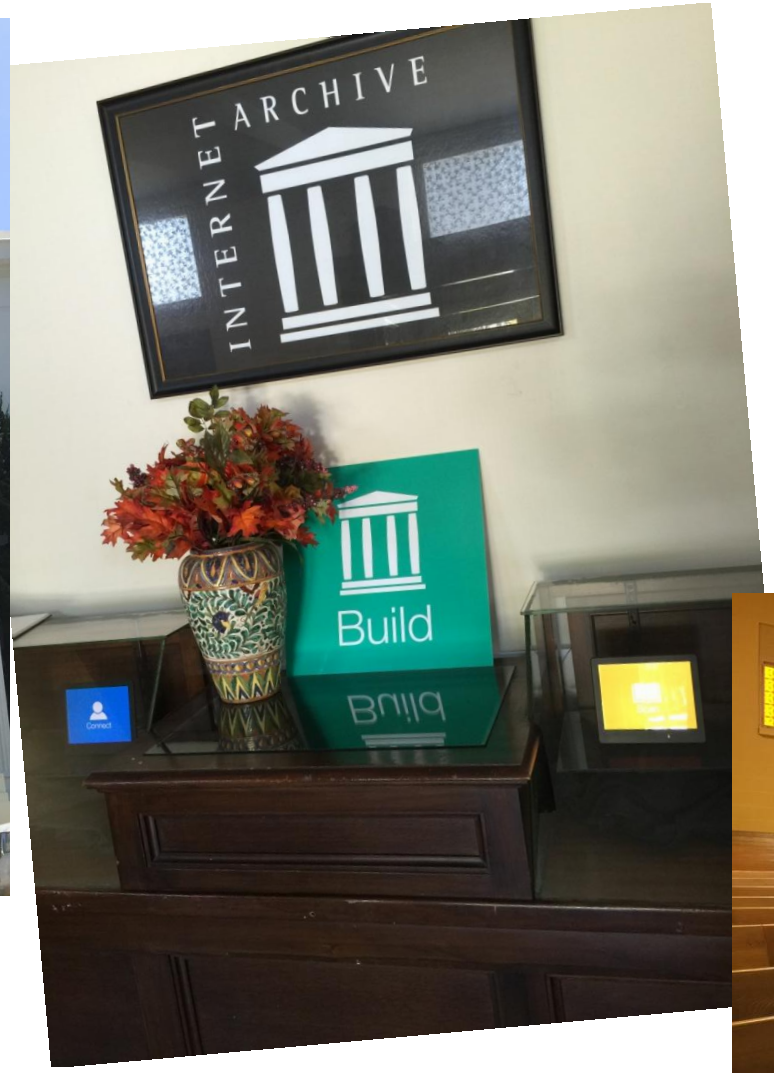
澳大利亚Pandora项目、瑞典Kulturarw项目，以及英国国家图书馆、法国国家图书馆等陆续启动网络信息采集

跨机构、跨国界的合作项目不断涌现，如2003年IIPC成立；英国网络存档联盟UKWARC成立；2004年Pandora项目已有10个澳大利亚州立图书馆加入；

国外网络信息保存的发展历程

2007年，国家图书馆加入了IIPC共同致力于网络信息保存工作。

该机构在网络信息保存系统架构、标准规范、元数据等方面建立了一系列技术规范，并资助其成员开发了从网络信息采集到提供访问服务的一系列高质量、易用的开源软件工具，包括采集工具（Heritrix、DeepArc、Smart Crawler）、资源保存工具（Web Curator Tool、NetarchiveSuite）、资源收藏存储工具（BnFArcTools）、访问工具（Wayback、NutchWAX、WERA、Xinq），为了方便和促进WARC格式的使用,还资助开发了WARC工具。


INTERNET ARCHIVE
Build

国外网络信息保存内容选择与规划

1.美国

美国国会图书馆

- 国内专题12个 重点关注政治、法律组织团体等
- 国外专题8个 关注热点地区的政治活动、重大宗教事件
- 保存内容类型包括博客、期刊杂志、音视频、电子信件等，在技术对较难采集的内容有所尝试

互联网档案馆（IA）

- “时光机”（Wayback Machine）项目为世界最大规模网络信息保存项目，内容包括各国大多数网址。
- “存档它”（Archive-It）项目公开的专题收藏有3007个（截止于2015年5月11日）。作为基于第三方的网络信息存档服务项目，存档内容由签订服务的合作机构决定。该项目共有超过300家合作机构。

国外网络信息保存内容选择与规划

2.英国

英国国家图书馆

- 受益于英国《非印刷品法定呈缴条例》，UKWARC项目保存全英国网络域名。
- 加强专题采集
- 长期维持“核心网站”清单。

英国国家档案馆

- “英国政府网络档案”
- 2013年12月起开始采集英国政府网站（详细内容可参看data.gov.uk）的全部数据。

国外网络信息保存内容选择与规划

3.澳大利亚

澳大利亚国家图书馆PADORA项目

- 11家会员机构；
- 在网络信息采集方面分工明确，每个成员设立自己的采集策略，澳大利亚国会图书馆负责采集具有全国性意义的网站和网页，地方图书馆采集具有地方意义的网站；
- 存档内容类型包括网页、网络出版物、影片、声音档案、各种多媒体动态形式的信息以及文本信息资源等，优先存档的是政府出版物、学术电子期刊。

澳大利亚国家图书馆与IA合作项目

- “存档它”项目合作，存档了世界顶级的关于亚洲-太平洋地区的50个专题网络档案；

国外网络信息保存内容选择与规划

4.法国、日本

法国国家图书馆

- 2010年起全面采集以.fr和.nc（New Caledonia）结尾的法国主域和所有在法国生产、出版发行或者发行商在法国的网上资源；
- 从IA获得1995至2004年的网络信息保存数据，2004至2009年合作存档，全面存档法国主域名；
- 结合主域、主题、事件主题存档。

日本国立国会图书馆

WARP项目

- 采集对象包括国家机关、独立行政法人、国立大学法人、银行等机构、地方公共团体等，涉密信息除外；
- 对于私立大学等私有机构或个人的具有国际性与文化性的主题网站，在获得所有者许可的基础上进行选择性采集，目前的采集率大约为6%。

国外项目数据采集

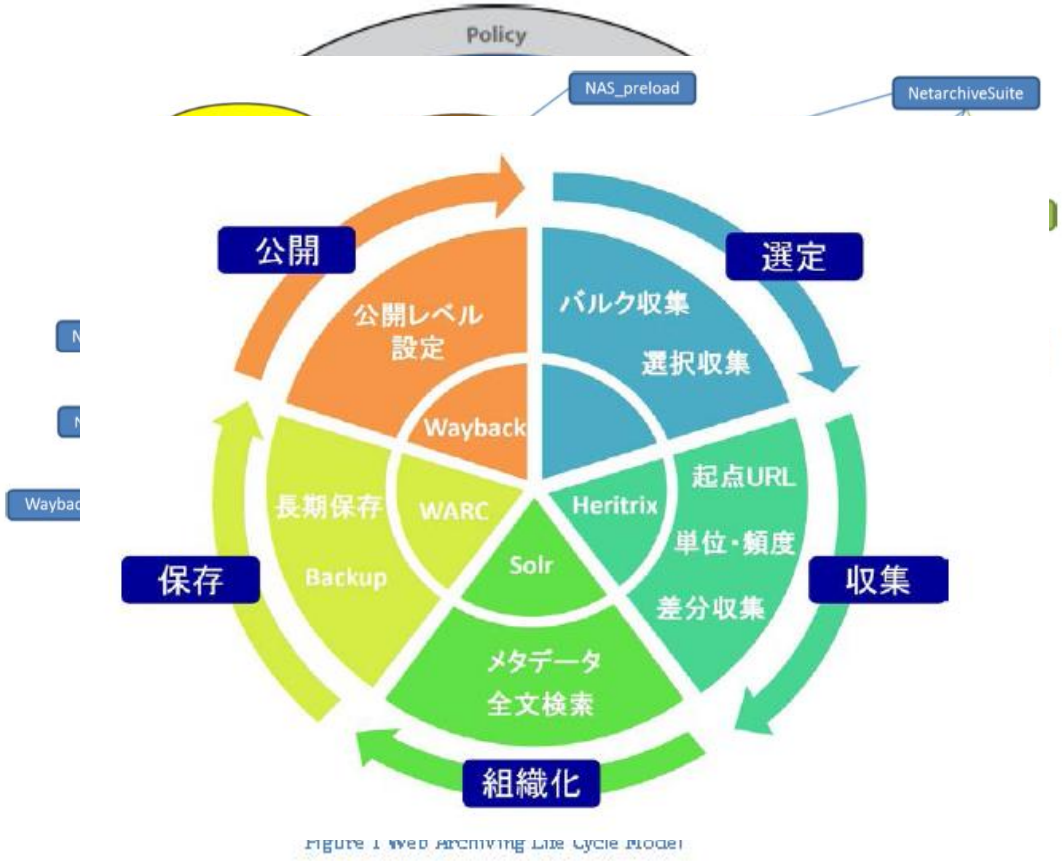
1.采集方式与采集频率

国家图书馆	采集方式	采集频率
法国	国家全域	每年1次
	选采	每周1次、每月1次、每年1-2次
英国	国家全域	每年1次
	选采	
挪威	国家全域	每年1次或2次
克罗地亚	国家全域	每年1次
瑞典	国家全域	每年2次
	选采140家报纸	每天1次
冰岛	国家全域	每年3次
	选采	至少每周1次
捷克	国家全域	每年1到2次
	选采	1500个网站每两个月1次
新加坡	国家全域	
	选采	每个季度1次
中国	政府全域（.gov)	每年1次
	选采	
日本	批量采集	每月1次
	选采	每年平均1-3次，每年4次
奥地利	国家全域	
西班牙	国家全域	
芬兰	国家全域	
加拿大	选采	政府网站半年1次
澳大利亚	选采	每月1次
荷兰	选采	
丹麦	批量采集	每年4次

国外项目数据采集

2.工作流程

法国国家图书馆网络档
IA 互联网档案馆项目工作流程
存项目工作流程
存档生命周期模式
日本国立国会图书馆
WARP项目工作流程



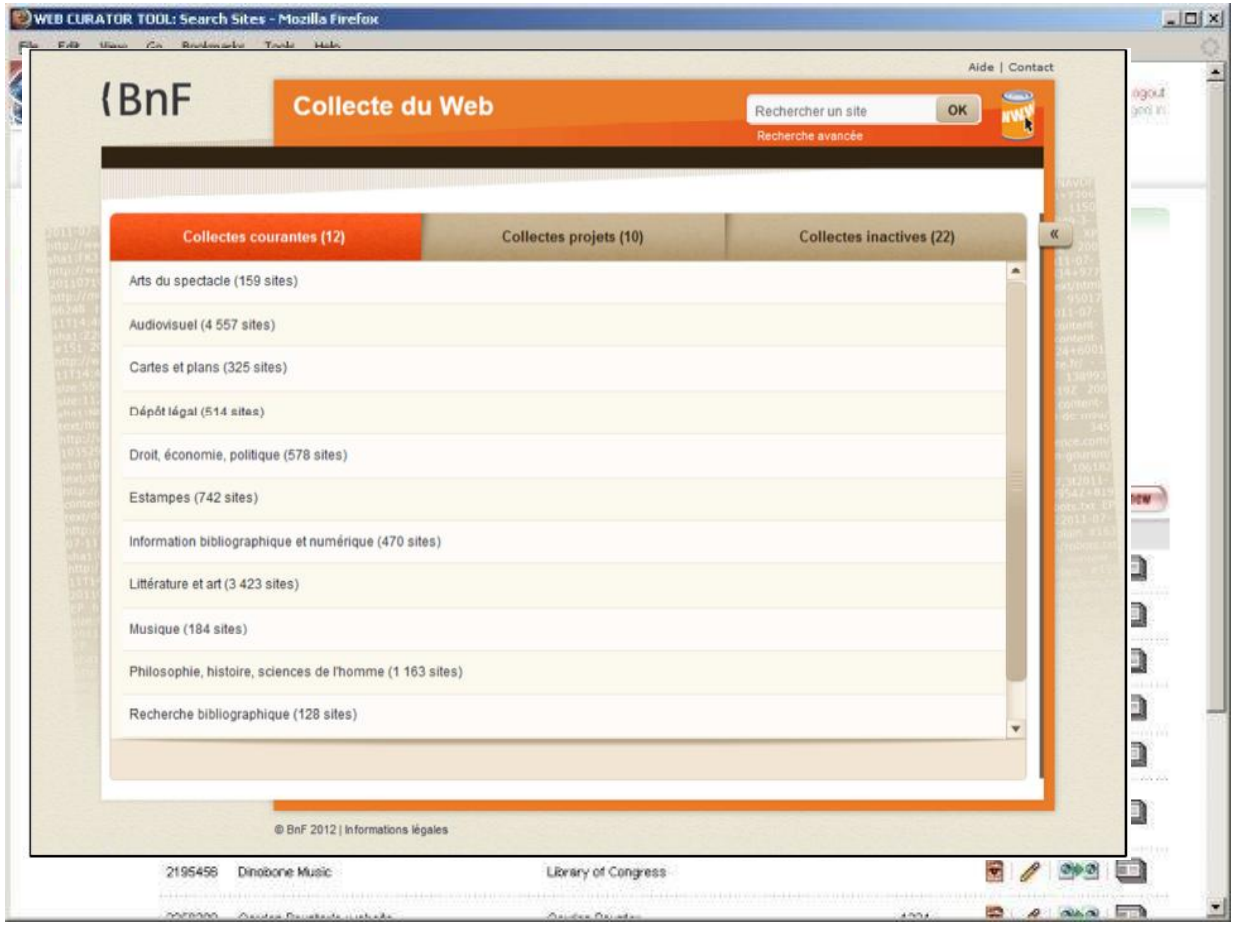
国外项目数据采集

2.采集软件

爬虫软件Heritrix

选择性采集工具BCWeb

采集管理工具WCT



国外项目数据管理与保存

1.存储设备

日本：
比特列保存 **RAID**
逻辑性保存
Migration
Emulation

IA
5
个
（
（**SAN**）和固态硬盘
（**SSD**）上

UK

国立国会図書館
National Diet Library

preserving uk websites

国外项目数据管理与保存

INTERNET ARCHIVE

WayBackMachine

保存网页超过**4000亿**个，数据总量将近**9PB**

Archives de l'internet

Les « archives de l'internet », photographies de l'internet français constituées par la BnF, peuvent être consultées en Bibliothèque de recherche.

保存**200亿**个网络资源对象，数据总量**450TB**

UK WEB ARCHIVE
preserving uk websites

每月约对**14000个**网站进行**70000次**存档

国外项目数据管理与保存

PANDORA
AUSTRALIA'S WEB ARCHIVE

完成了对**10万**个主题的保存工作，保存的文件数达到了**9.5亿**个，数据量共**31.93TB**

Domain Harvest	Unique files	Hosts	Size (TB)
2005	185 m	811,523	6.69
2006	596 m	1,046,038	19.04
2007	516 m	1,247,614	18.47
2008	1 billion	3,038,658	34.55
2009	756 m	1,074,645	24.28
2011	660 m	1,346,549	30.71
2012	1 billion	1,467,158	41.88
2013	660 m	1,690,232	29.17
2014	953 m	7,046,168	31.93

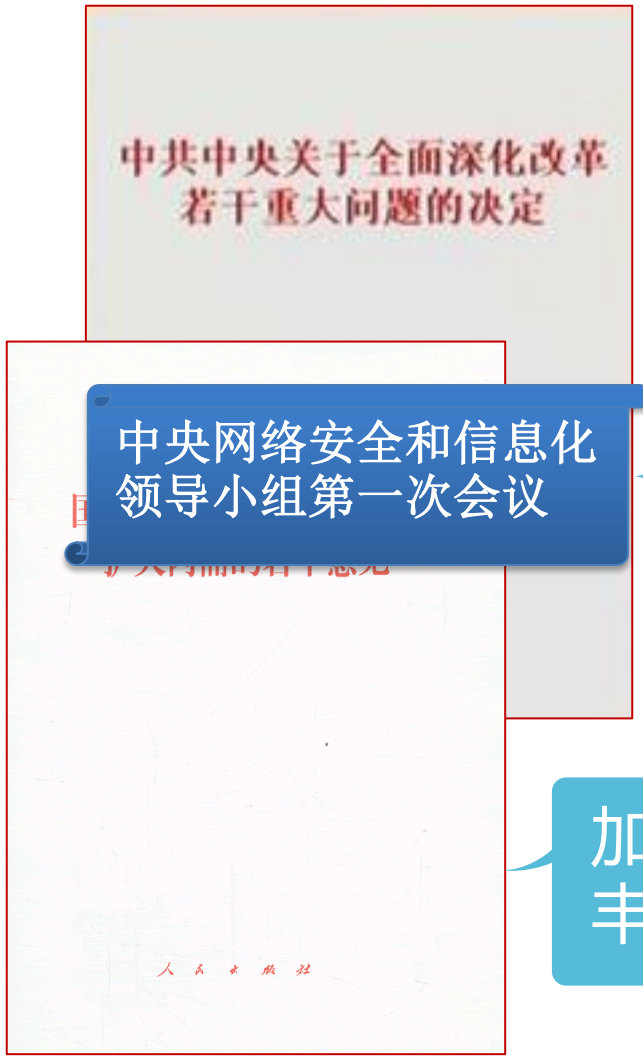
国外项目发布与服务


PANDORA
AUSTRALIA'S WEB ARCHIVE
Search Trove
Search Help
Provided by:
BRITISH LIBRARY
Home
About PANDORA
Brief overview
History
Policy and practice
Selection guidelines
Manuals
PANDAS
Staff papers
Digital preservation
FAQ
Key Studies
Legal Deposit
Blog
Partners
Notification form
Services
Statistics
Contact us
Other archives
Disclaimers
NLA home page
Home > About PANDORA > Selection Guidelines
SELECTION GUIDELINES
PANDORA is a selective archive. The National Library and its partners do not attempt to collect all Australia
significance and to have long-term research value.
Some national libraries, such as Australia and Canada, have established selective archives, and some, such
snapshots of their countries'entire web domains. There are advantages and disadvantages to both approache
periodic snapshots of the entire domain. For a discussion of the advantages and disadvantages of both appr
Publications.
Each of the PANDORA participating agencies selects titles for the Archive according to its own selection gui
national significance, while the State libraries aim to archive those of State and regional significance. Screer
Australian War Memorial archives sites relating to Australian military history; and the Australian Institute of A
our Indigenous peoples.
PANDORA Participating Agencies' Selection Guidelines
Australian Institute of Aboriginal and Torres Strait Islander Studies
Australian War Memorial
National Film and Sound Archive (ScreenSound Australia)
National Library of Australia
Northern Territory Library and Information Service
State Library of New South Wales
State Library of Queensland
State Library of South Australia

4

我国网络信息保存与利用
——蓄势待发

1、良好的政策环境



中央网络安全和信息化领导小组第一次会议

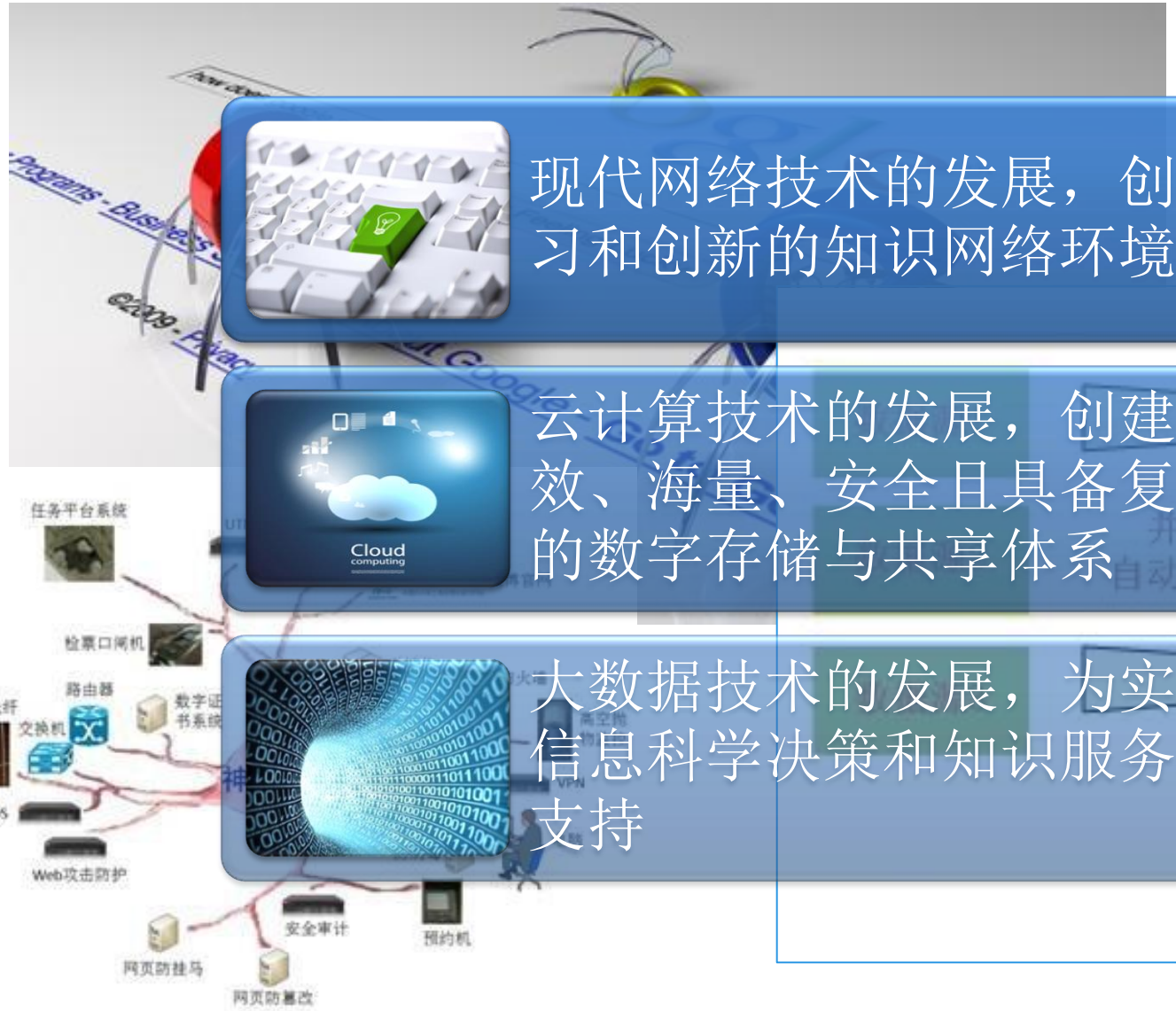
面对互联网技术和应用飞速发展，现行管理体制存在明显弊端，多头管理、职能交叉、权责不一、效率不高。

习近平发表重要讲话：网络安全和信息化是事关国家安全和国家发展、要努力把我国建设成为网络强国。

习近平强调,网络信息是跨国界流动的,信息流引领技术流、资金流、人才流,信息资源日益成为重要生产要素和社会财富,信息掌握的多寡成为国家软实力和竞争力的重要标志。

加强信息基础设施建设，加快信息产业优化升级，大力丰富信息消费内容，提高信息网络安全保障能力

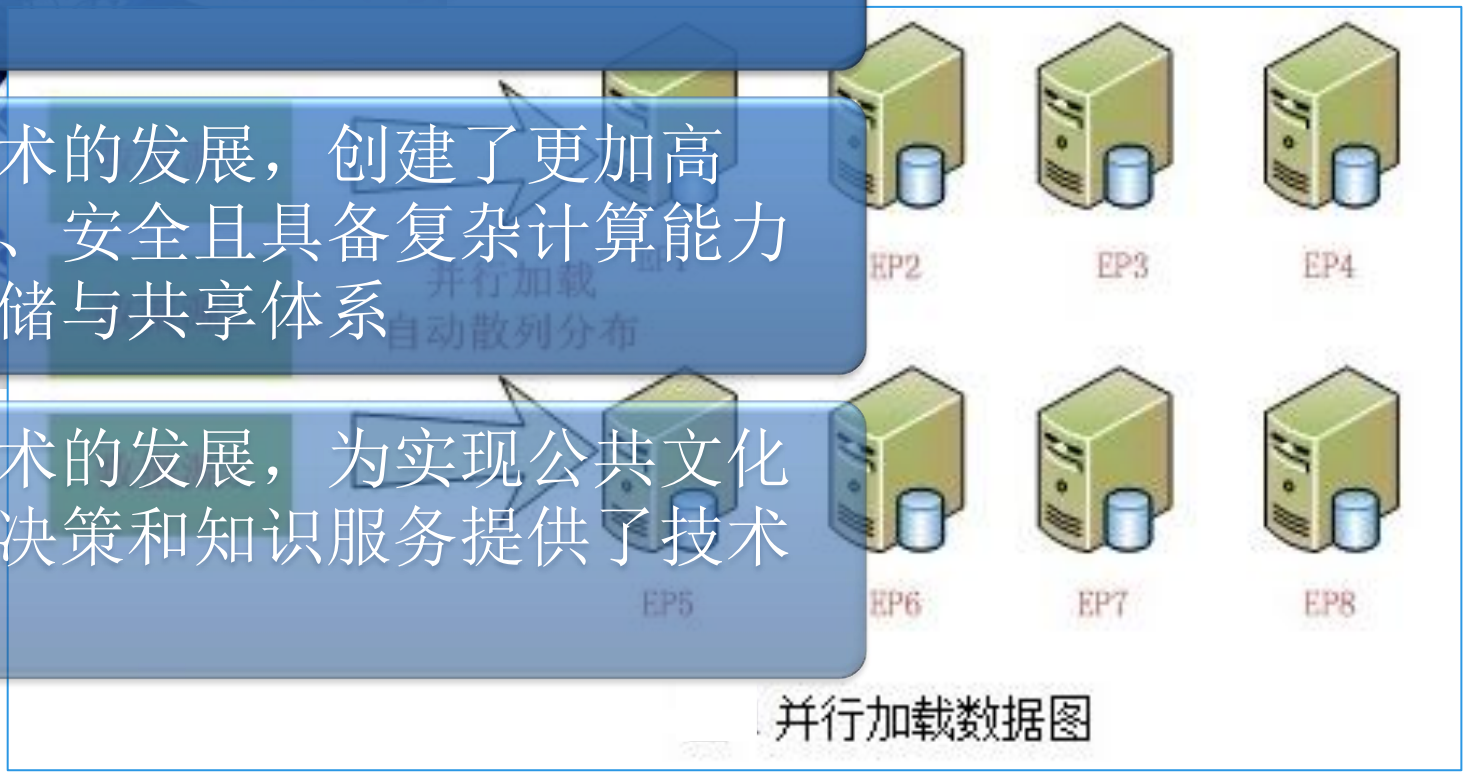
2、坚实的技术支撑



现代网络技术的发展，创建了支持学习和创新的知识网络环境

云计算技术的发展，创建了更加高效、海量、安全且具备复杂计算能力的数字存储与共享体系

大数据技术的发展，为实现公共文化信息科学决策和知识服务提供了技术支持



并行加载数据图

3、广泛的实践基础

在国家973和985项目支持下，北京大学网络实验室开发建设的中国网页历史信息存储与展示系统，从2002年1月18日上线运行。

截止2014年7月，Web Infomall保存网页**89亿**，总容量**73TB**

Web Infomall尤其重视网络信息保存相关方面的技术研发，强调拥有自主知识产权，为我国的网络信息的发展奠定了良好的技术基础

5

国家图书馆网络信息保存工作进展

浅层网络信息整合

2003年初，中国国家图书馆开展了**网络信息资源采集与保存试验项目（WICP）**

2009年，国家图书馆成立了**国家图书馆互联网信息保存保护中心**，并于2012 年开通网站服务。国家图书馆互联网信息保存保护中心是中国国家图书馆成立的致力于中国互联网信息资源长期保存和保护的机构，它的成立可以说是中国的互联网保护事业的里程碑，开创了中国互联网保护工作统筹规划、合作共建的新局面。

研究网络资源评估体系，完善资源采选方案

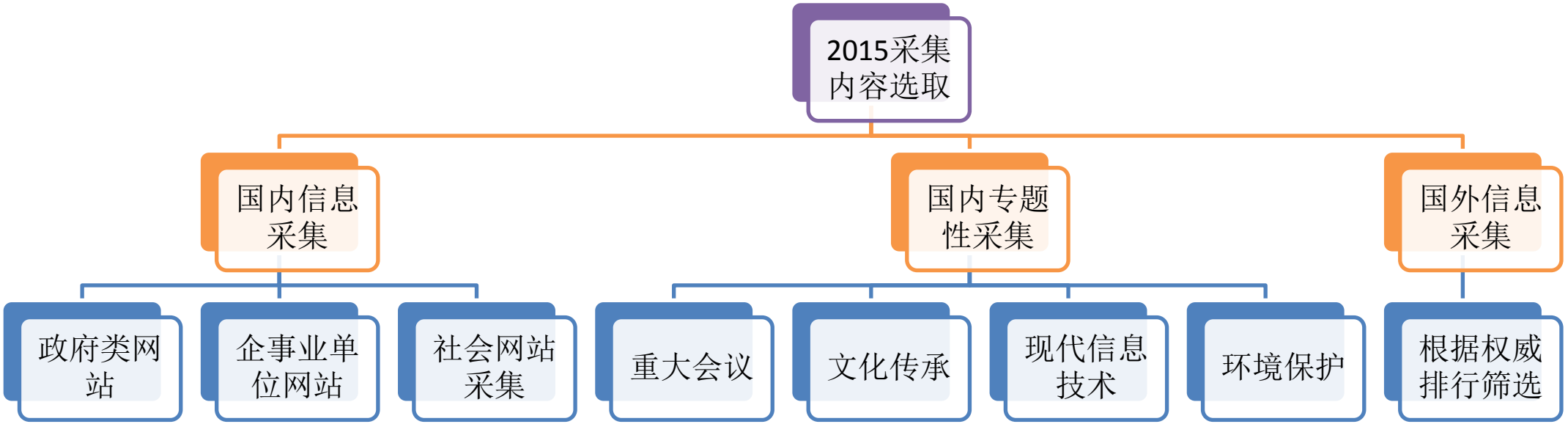
确定层次分明的网络信息采集任务，形成布局合理、结构优化的资源体系

对现有各种最新版本的采集软件进行调研评测

加强技术调研与研发，深入分析国内外主要项目系统架构、存储机制和关键技术等做，规划海量数据存储方案

加强宣传推广力度，扩大工作的社会影响力，形成多层级、分布式的资源组织与服务体系

采集内容选取方案

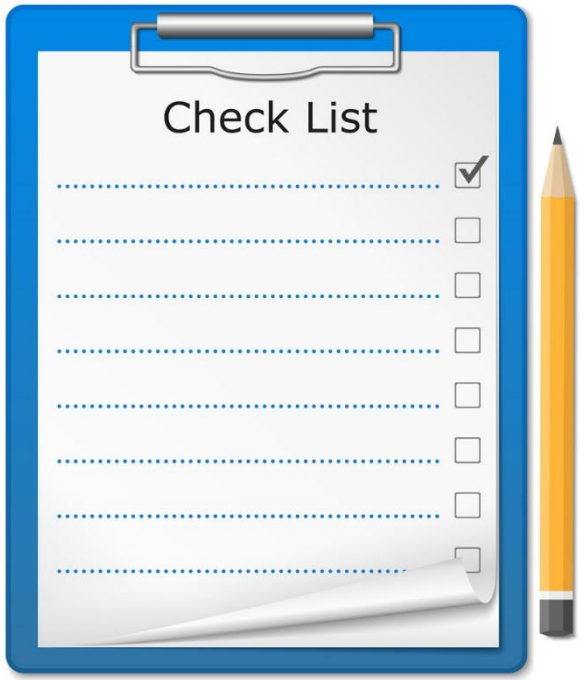


（一）采集策略

1、对采集清单权威性进行确认

- 前期规划共同完成采集内容列表的选取与制定

- 初步筛选了涉及**13个行业**以及门户网站的**1464个域名**清单和**82个中央级国家机关网站**清单，经过百度对网页综合指标测评，排名也是靠前的，说明网页质量和权威性有保障。

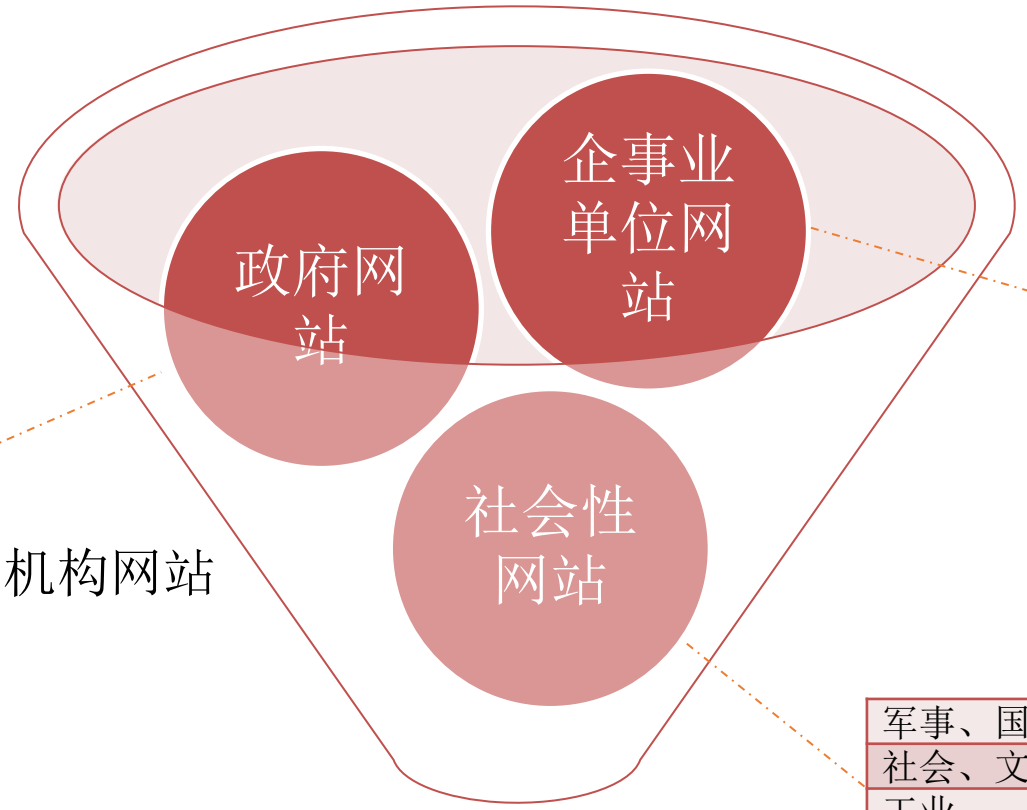


2、确定采集技术策略

- 国际性网络信息保存项目的采集方式主要为：
  - 整站采集(对某网站域名下的所有网页进行采集)、
  - 选择性采集(只选择某网站特定的专题事件、特定页面或资源进行采集)，
  - 混合型采集(兼顾整站采集和选择性采集两种方式)。

- 根据国家数字图书馆网络信息保存工作的发展规划和服务需要，结合现有存储条件及技术应用水平，本项目拟采取**混合型采集方式**

2、确定采集技术策略



限于我国中央级国家机构网站

国有重点企业及中央企业网站
重点直属事业单位网站
民主党派网站
重要社会组织事业单位网站

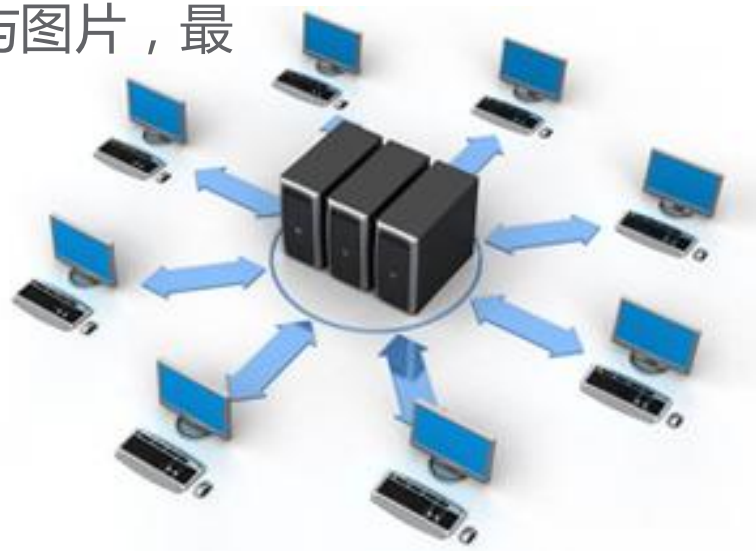
军事、国防	经济	教育、科研	体育
社会、文化	文学、艺术	医药、卫生	农业
工业	交通	航空、航天	环境、安全
宗教			

国内网站

2、确定采集技术策略

- 目前商业化搜索引擎采集模式：采集内容以静态网页为主，通过主页对该站静态页面进行链接
  - √ **优点**：此数据提交后不需进一步加工即可直接在网站上发布和回放，资源发布的效率较高、存储成本较低；
  - × **缺点**：对象数据（如图片）以链接地址进行显示，回放需链接互联网且存在对象数据链接地址过期失效的问题；

- 我们参考借鉴国外主流的网络信息保存项目，采集内容包括文本与图片，最终实现的是网络信息基于本地服务器的原始回放。



Baidu 百度

- 百度是最大的中文搜索引擎、最大的中文网站，拥有全球三大网页库之一
- 以**“让人们最平等、便捷地获取信息，找到所求”**为使命
- 致力于提供“**简单，可依赖**”的互联网搜索产品及服务

3、数据交付和临时存储

- 鉴于数据量庞大，且我馆带宽有限，协商采用我们提供临时存储设备的方式，每周中转一次数据，保证及时验收、整理和发布。

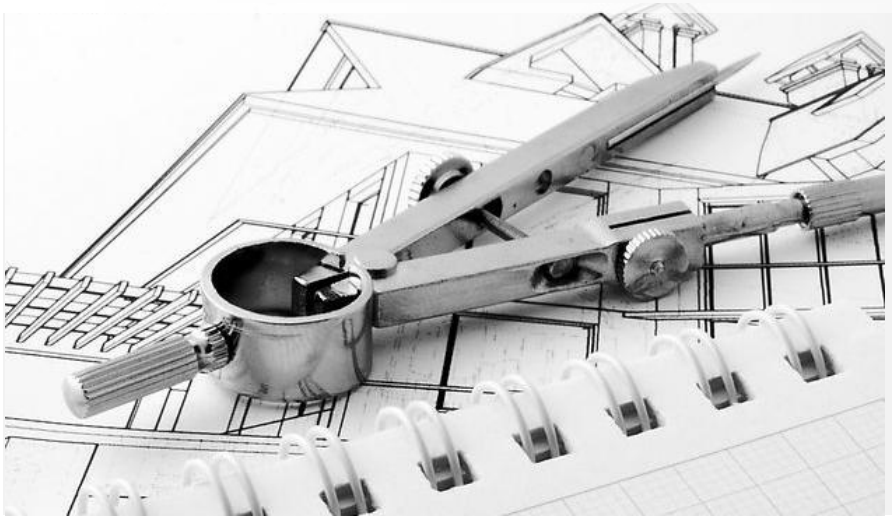
1、合作目标

实现学术资源的共建、共知和共享，在信息资源建设和保存、知识发现与获取等方面进行多方位交流与合作；

充分发挥百度海量用户的入口优势以及国图海量资源的优势，形成图书馆公共文化资源与百度互联网搜索引擎技术的深度融合；

更大范围内实现资源服务高效便捷、传播快捷的有益探索，将公共数字文化服务惠及全民的理念落到实处。

2、合作范围



落实互联网+环境下国家战略部署，充分发挥双方在国家信息化建设和现代公共文化服务体系构建中的重要作用，

依托国家数字图书馆资源优势，依托百度技术创新优势和数据处理能力，共同推进公共数字文化建设和发展,进而深入推进重点惠民工程建设。

秉承公共文化服务公益性、基本性、均等性、便利性原则,国家数字图书馆将与百度同创共建公共文化数字产品，提升公共数字文化服务能力。

4、确定今年采集数量

- 我馆结合网络资源增长趋势及采集、存储条件的相应发展规划，本年度由百度完成：
  - **1546个**站点
  - **5万次**采集任务
  - 数据总量不少于**80TB**
  - **8亿个**网页

采集工具：Heritrix

Heritrix是目前使用最广泛的网站爬虫（crawler）。它基于java平台开发，并进行了开源处理。当前最新的Heritrix版本为互联网档案馆（IA）与北欧多个国家图书馆开发并于2014年1月更新的Heritrix3.2.0版。

访问工具：Wayback

Wayback软件是由IA开发，是目前Web Archive中应用最为广泛的检索和浏览工具。Wayback是一款开源软件，允许用户检索资源集来定位文档，浏览服务允许用户通过web浏览器查看资源集中的存档文档，多种访问方式（存档URL模式、协议URL模式、域名前缀模式）供用户选择实现网站重现。

（二）数据采集

国家图书馆网络信息采集数据量

数据类型	年份	采集政府网站数量	采集政府网站数据（TB）	采集专题网站数量	采集专题网站数据(TB)	年度合计数据（TB）
采集数据	2005	19968	0.25	193	0.20	0.45
	2006	11317	0.89	614	0.22	1.11
	2007	0	0	469	0.45	0.45
	2008	50000	7.52	326	3.18	10.70
	2009	35	3.52	225	0.32	3.84
	2010	75	1.38	443	0.80	2.18
	2011	0	0	314	0.37	0.37
	2012	20138	10.99	268	0.77	11.76
	2013	20738	10.07	336	0.31	10.38
	2014	19164	15.723	387	0.449	16.172
合计：		**141435**	**50.343**	**3575**	**7.069**	**57.412**

（二）数据采集

资源保存格式标准：Warc

在国际标准“ISO28500——Information and documentation-WARC file format”的基础上制定符合我国国情和中文语境的网络资源采集与存档标准，使存档文件能简单并安全地承载网络资源的大量数据对象，以便进行存储、管理和交换。

存储设备

当前，国家图书馆网络信息采集资源的存储空间约为110T。其中存域网空间15T，另有7台服务器既用作存储也用作发布。

（三）数据保存

国家图书馆互联网信息保存保护中心网站

（三）数据发布

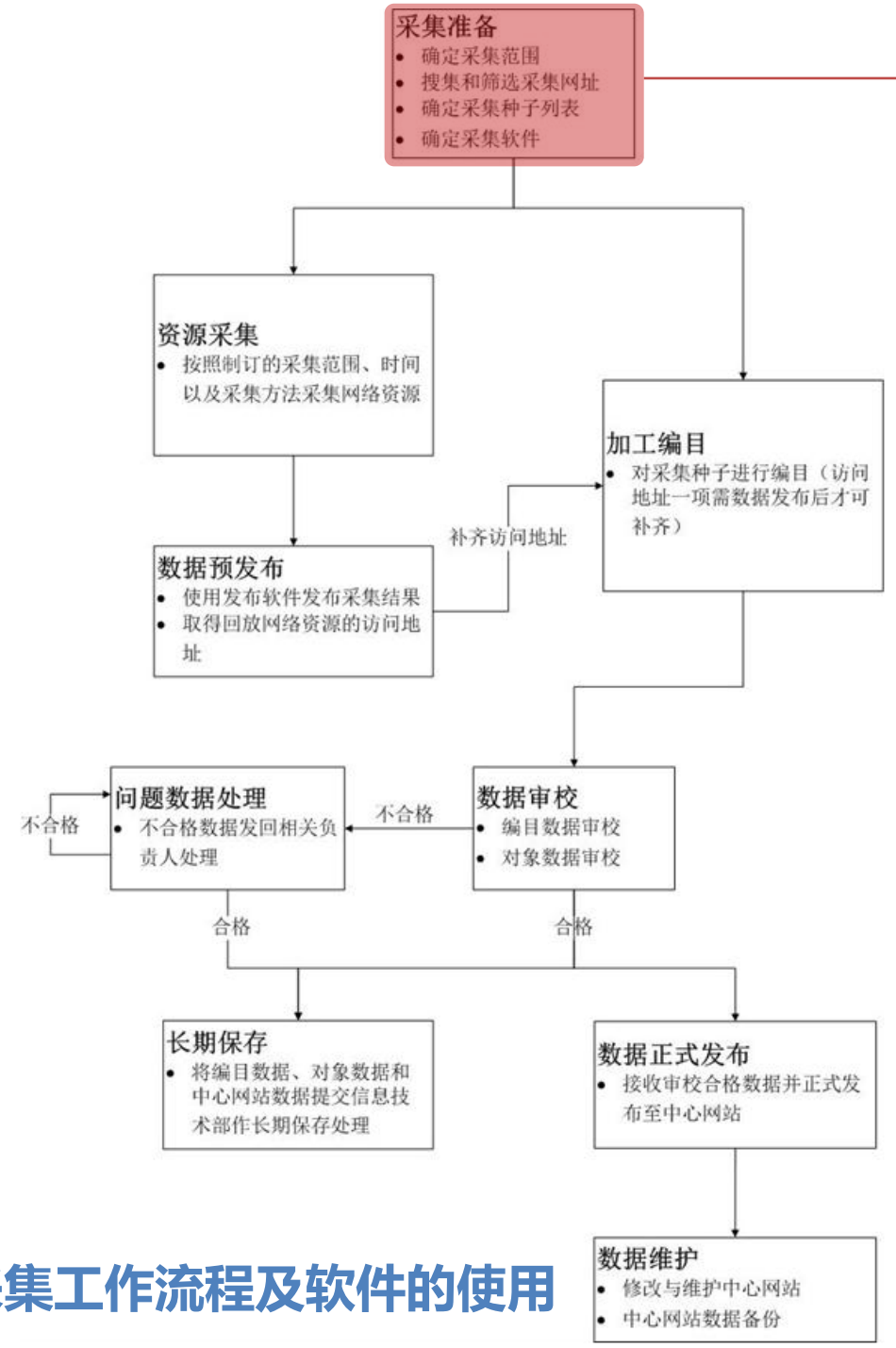
国家图书馆互联网信息保存保护中心网站

◆当前正在进行改版，将由局域网访问转为实现互联网公开访问；

◆结合采集内容和数量，搭建以网站回放和专题揭示为主的栏目结构，开展网页资源的深层次挖掘和可视化展示;

◆通过主题、时间、国别等进行多维分类导引，完成当年所有采集数据的索引和发布服务,以多角度、多种方式展示发布资源，深度整合网络资源采集和发布的工作成果。

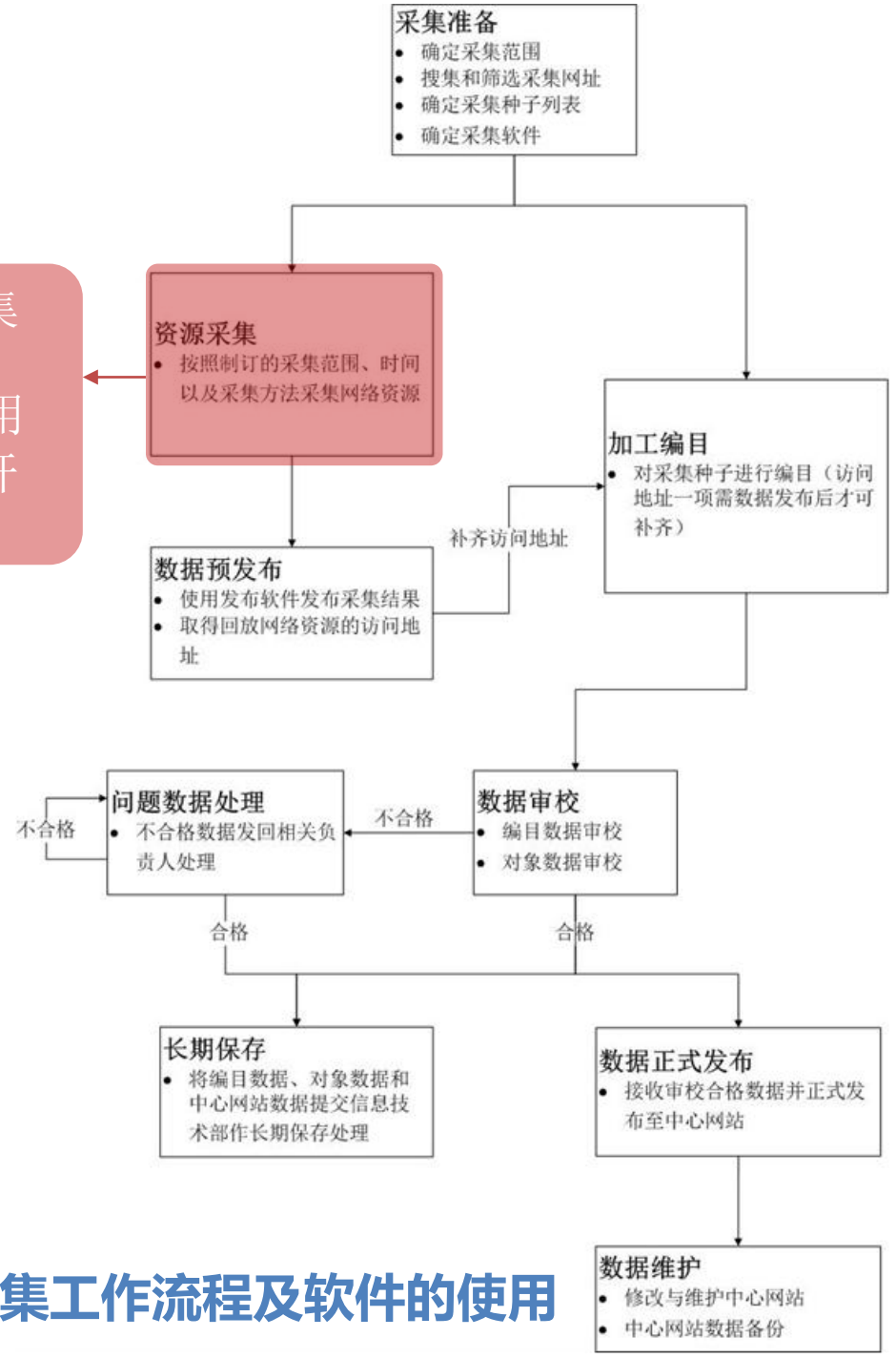
（三）数据发布



根据采选原则，对网络资源进行梳理，确定采集范围，完成采集目标种子网址（URL地址）的搜集和筛选，整理成网络采集的资源种子网址列表。

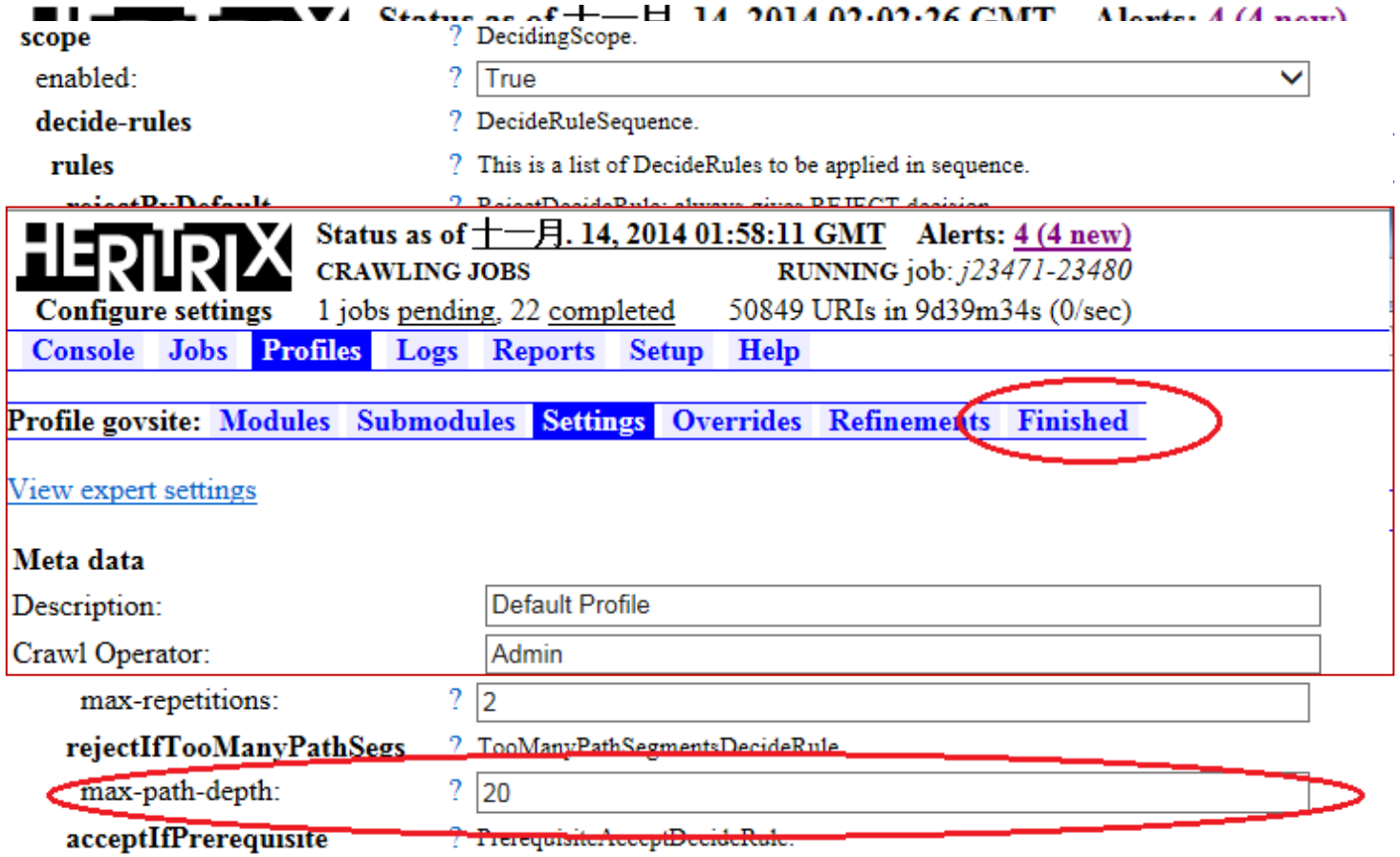
（四）开源系统的网络资源采集工作流程及软件的使用

根据采集要求，利用网络采集软件，对种子网址进行采集。两个软件：**Heritrix**（国际通用的开源软件）和国家图书馆开发的**网页资源获取系统**



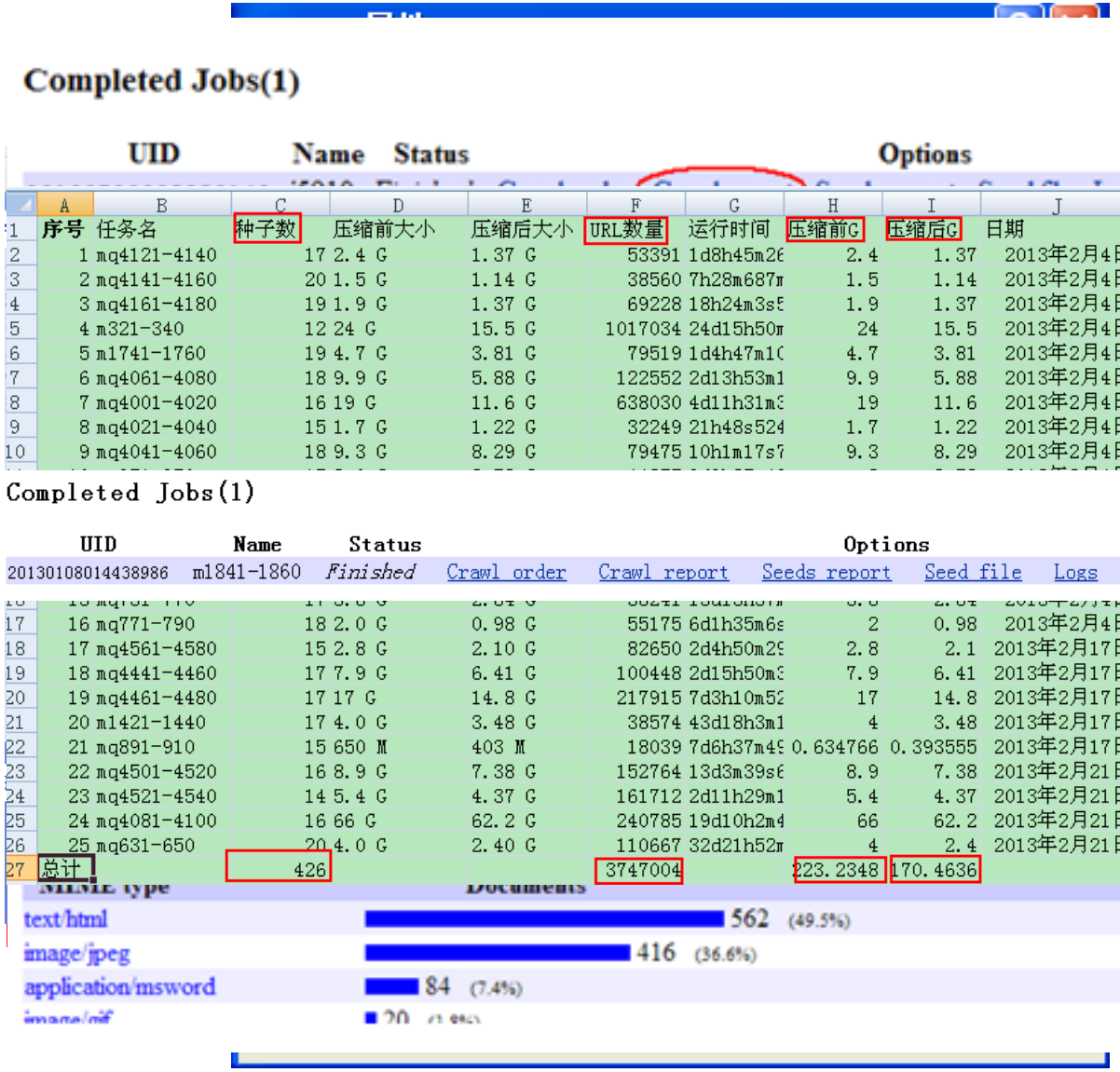
（四）开源系统的网络资源采集工作流程及软件的使用

Heritrix采集流程



（四）开源系统的网络资源采集工作流程及软件的使用

Heritrix采集任务实例



Completed Jobs(1)

UID	Name	Status	Options

序号	任务名	种子数	压缩前大小	压缩后大小	URL数量	运行时间	压缩前G	压缩后G	日期	服务器IP	是否转移
1	mq4121-4140	17	2.4 G	1.37 G	53391	1d8h45m2	2.4	1.37	2013年2月4日	125.5	已转移
2	mq4141-4160	20	1.5 G	1.14 G	38560	7h28m687	1.5	1.14	2013年2月4日	125.5	已转移
3	mq4161-4180	19	1.9 G	1.37 G	69228	18h24m3s	1.9	1.37	2013年2月4日	125.5	已转移
4	m321-340	12	24 G	15.5 G	1017034	24d15h50	24	15.5	2013年2月4日	125.5	已转移
5	m1741-1760	19	4.7 G	3.81 G	79519	1d4h47m1	4.7	3.81	2013年2月4日	125.5	已转移
6	mq4061-4080	18	9.9 G	5.88 G	122552	2d13h53m1	9.9	5.88	2013年2月4日	125.5	已转移
7	mq4001-4020	16	19 G	11.6 G	638030	4d11h31m3	19	11.6	2013年2月4日	125.5	已转移
8	mq4021-4040	15	1.7 G	1.22 G	32249	21h48s524	1.7	1.22	2013年2月4日	125.5	已转移
9	mq4041-4060	18	9.3 G	8.29 G	79475	10h1m17s7	9.3	8.29	2013年2月4日	125.5	已转移

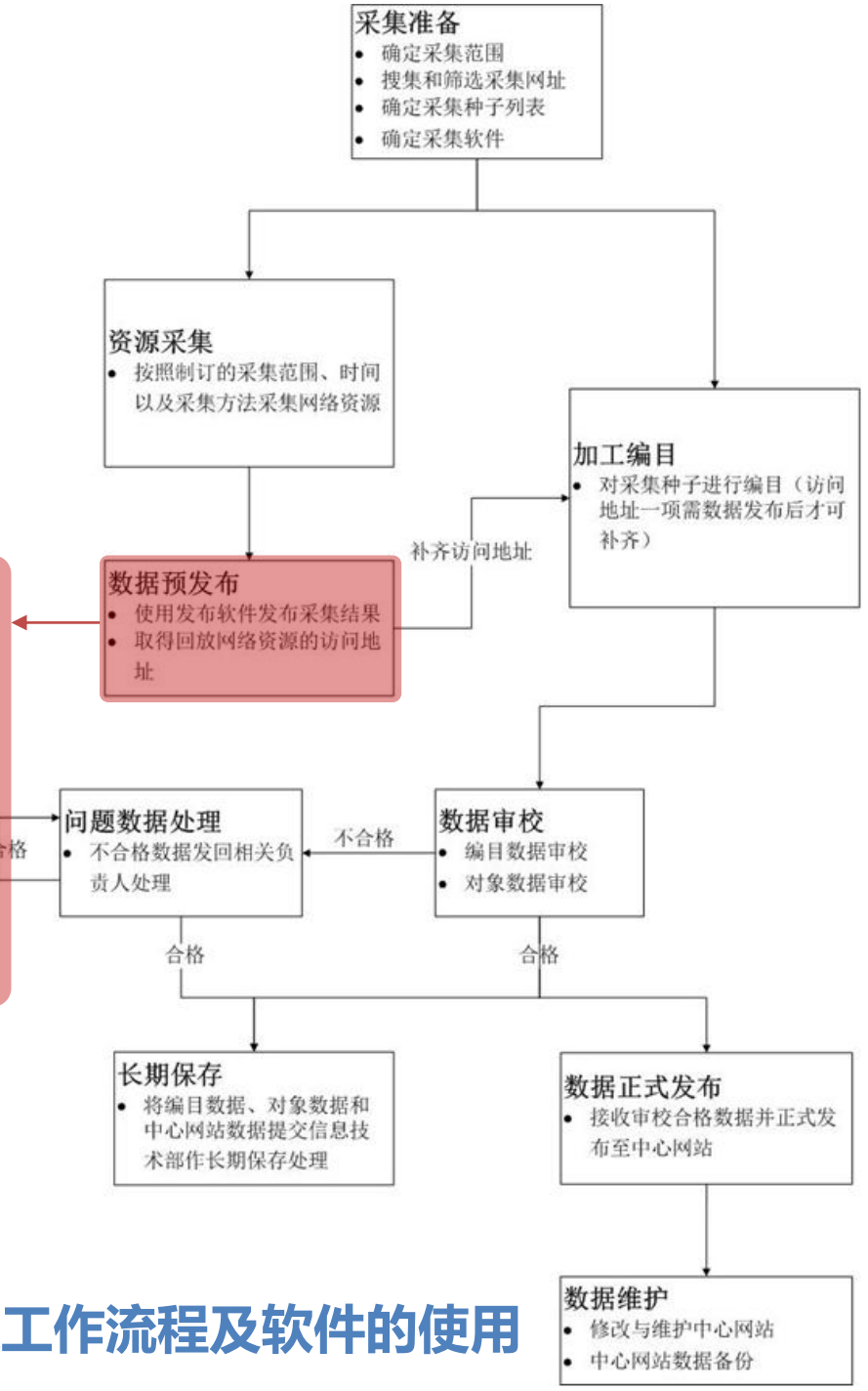
Completed Jobs(1)

UID	Name	Status	Options					
20130108014438986	m1841-1860	*Finished*	Crawl order	Crawl report	Seeds report	Seed file	Logs	Journal	Delete

16	mq771-790	18	2.0 G	0.98 G	55175	6d1h35m6s	2	0.98	2013年2月4日	180.132	已转移
17	mq4561-4580	15	2.8 G	2.10 G	82650	2d4h50m2s	2.8	2.1	2013年2月17日	125.5	已转移
18	mq4441-4460	17	7.9 G	6.41 G	100448	2d15h50m3	7.9	6.41	2013年2月17日	125.5	已转移
19	mq4461-4480	17	17 G	14.8 G	217915	7d3h10m52	17	14.8	2013年2月17日	125.5	已转移
20	m1421-1440	17	4.0 G	3.48 G	38574	43d18h3m1	4	3.48	2013年2月17日	180.1501i	已转移
21	mq891-910	15	650 M	403 M	18039	7d6h37m4s	0.634766	0.393555	2013年2月17日	180.133	已转移
22	mq4501-4520	16	8.9 G	7.38 G	152764	13d3m39s6	8.9	7.38	2013年2月21日	125.5	已转移
23	mq4521-4540	14	5.4 G	4.37 G	161712	2d11h29m1	5.4	4.37	2013年2月21日	125.5	已转移
24	mq4081-4100	16	66 G	62.2 G	240785	19d10h2m4	66	62.2	2013年2月21日	125.5	已转移
25	mq631-650	20	4.0 G	2.40 G	110667	32d21h52m	4	2.4	2013年2月21日	180.132	已转移
总计		426			3747004		223.2348	170.4636			

MIME type	Documents	Data
text/html	562 (49.5%)	59 MB
image/jpeg	416 (36.6%)	195 MB
application/msword	84 (7.4%)	12 MB
image/gif	20 (1.8%)	96 KB

（四）开源系统的网络资源采集工作流程及软件的使用



通过资源回放软件为归档数据建立索引，实现类别划分和访问控制，完成预发布。

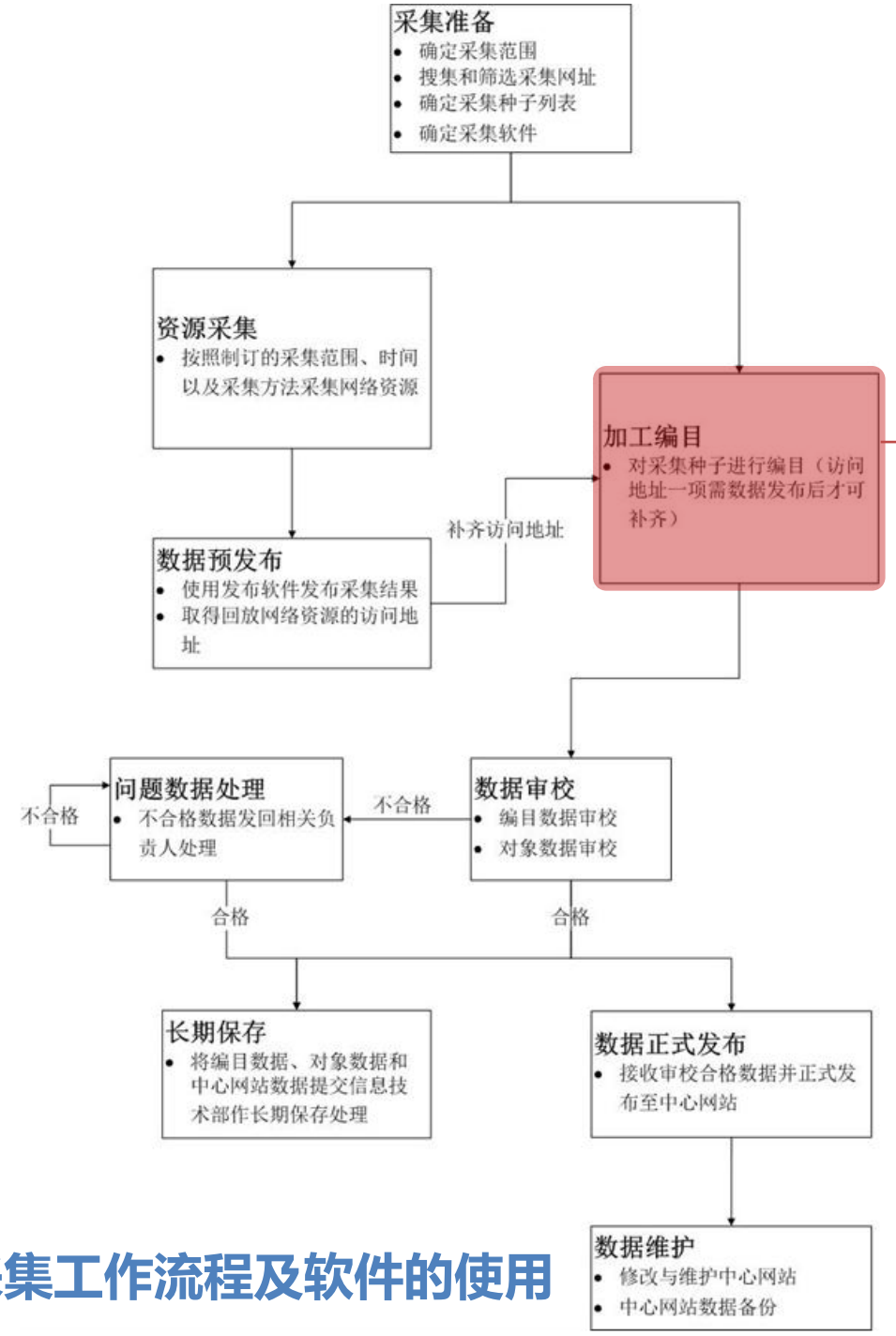
两个软件：**Wayback**（IIPC推荐的开源软件）和国家图书馆开发的**网页资源获取系统**。

（四）开源系统的网络资源采集工作流程及软件的使用

Wayback软件索引配置

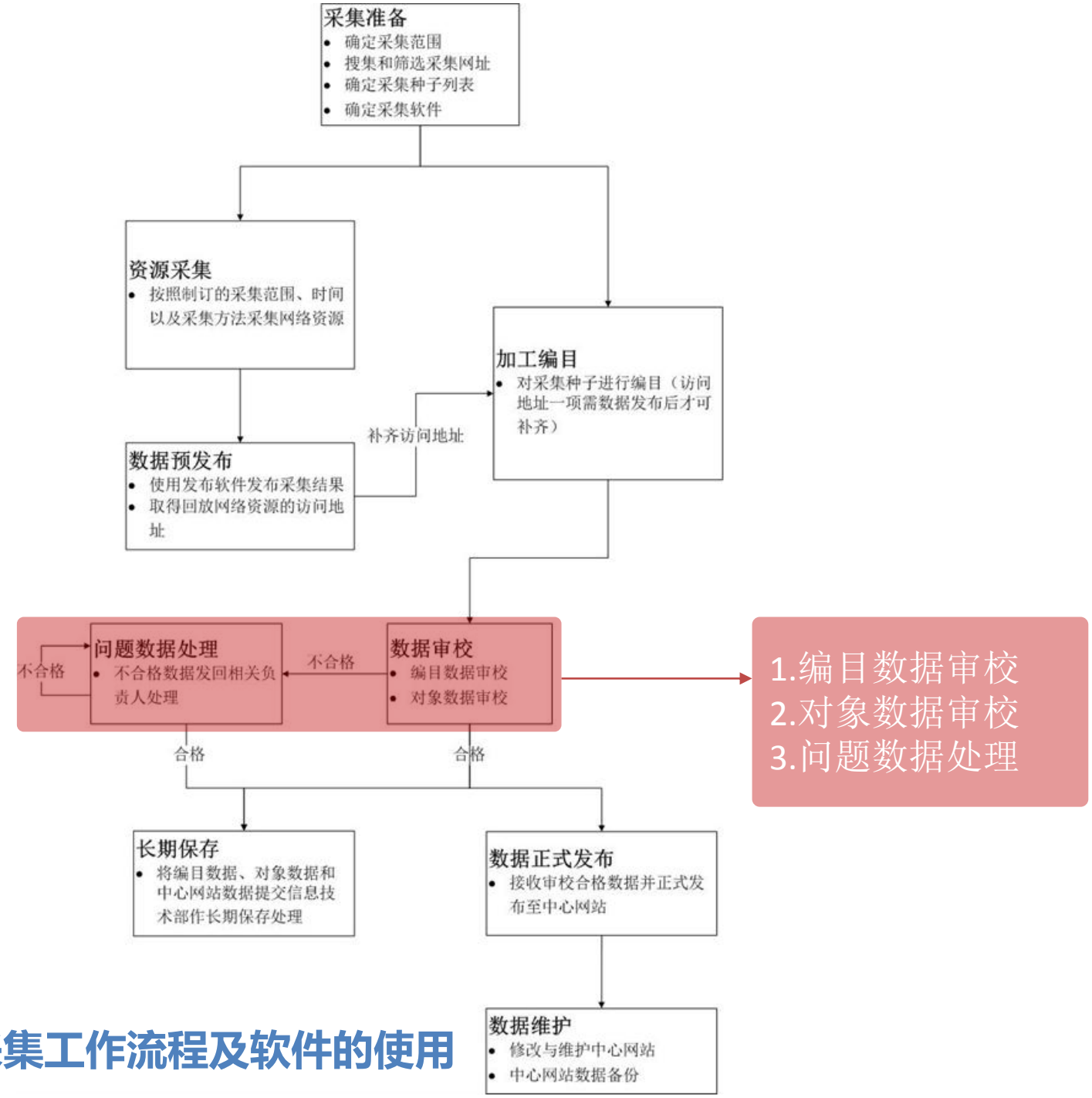
```
<bean id="datadirs2006" class="org.springframework.beans.factory.config.ListFactoryBean">
  <property name="sourceList">
    <list>
      <bean class="org.archive.wayback.resourcestore.resourcefile.DirectoryResourceFileSource">

<bean id="localbdbcollection2006" class="org.archive.wayback.webapp.WaybackCollection">

  <property name="resourceStore" ref="localresourcestore2006" />

  <property name="resourceIndex" ref="localbdbresourceindex2006"/>

  <property name="shutdownables">
    <list>
      <!-- This thread notices new files appearing in your resourcefilesources -->
      <bean id="resourcefilesourceupdater2006" class="org.archive.wayback.resourcestore.re
        <property name="target" value="${wayback.basedir2006}/file-db/incoming" />
        <property name="interval" value="100000" />
        <property name="sources" ref="datadirs2006" />
      </bean>

      <!-- This thread updates the location db with updates from resourcefilesourceupdater
      <bean id="resourcefilelocationdbupdater2006" class="org.archive.wayback.resourcestore
        <property name="interval" value="1000" />
        <property name="db" ref="resourcefilelocationdb2006" />
        <property name="incomingDir" value="${wayback.basedir2006}/file-db/incoming" />
        <property name="stateDir" value="${wayback.basedir2006}/file-db/state" />
      </bean>
        <property name="name" value="files200704" />
        <property name="prefix" value="${wayback.basewarcdir}/2007_chang_e" />
        <property name="recurse" value="false" />
      </bean>
      <bean class="org.archive.wayback.query.Renderer">
```

（四）开源系统的网络资源采集工作流程及软件的使用

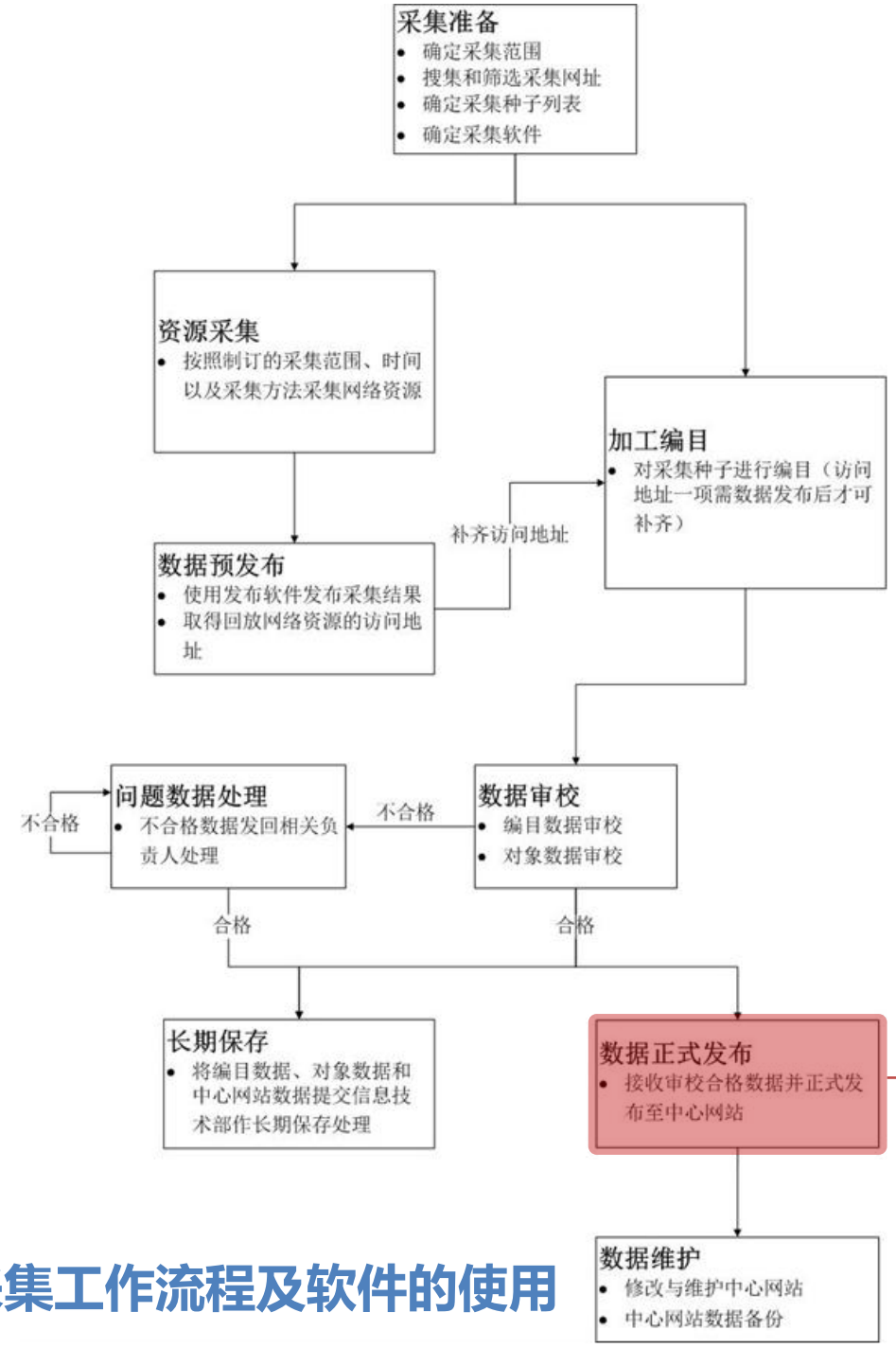


照相关著录规则（**政府及相关机构存档网站元数据著录规则**和**专题存档元数据著录规则**）对采集到的政府及专题存档资源进行编目，将编目信息制作成**excel表**（该表中访问地址一项需要在数据发布后才能取得补齐）

（四）开源系统的网络资源采集工作流程及软件的使用

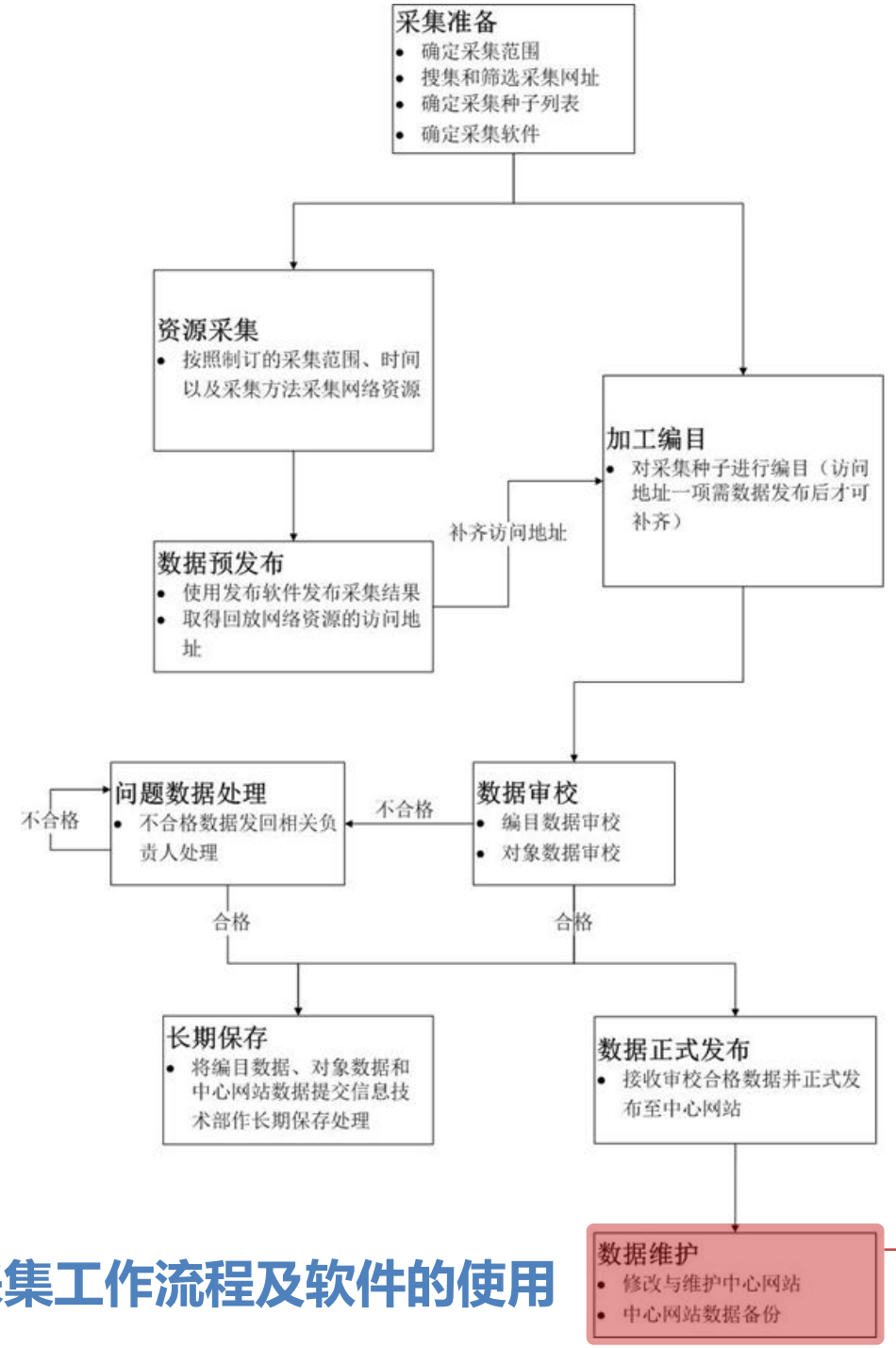


（四）开源系统的网络资源采集工作流程及软件的使用



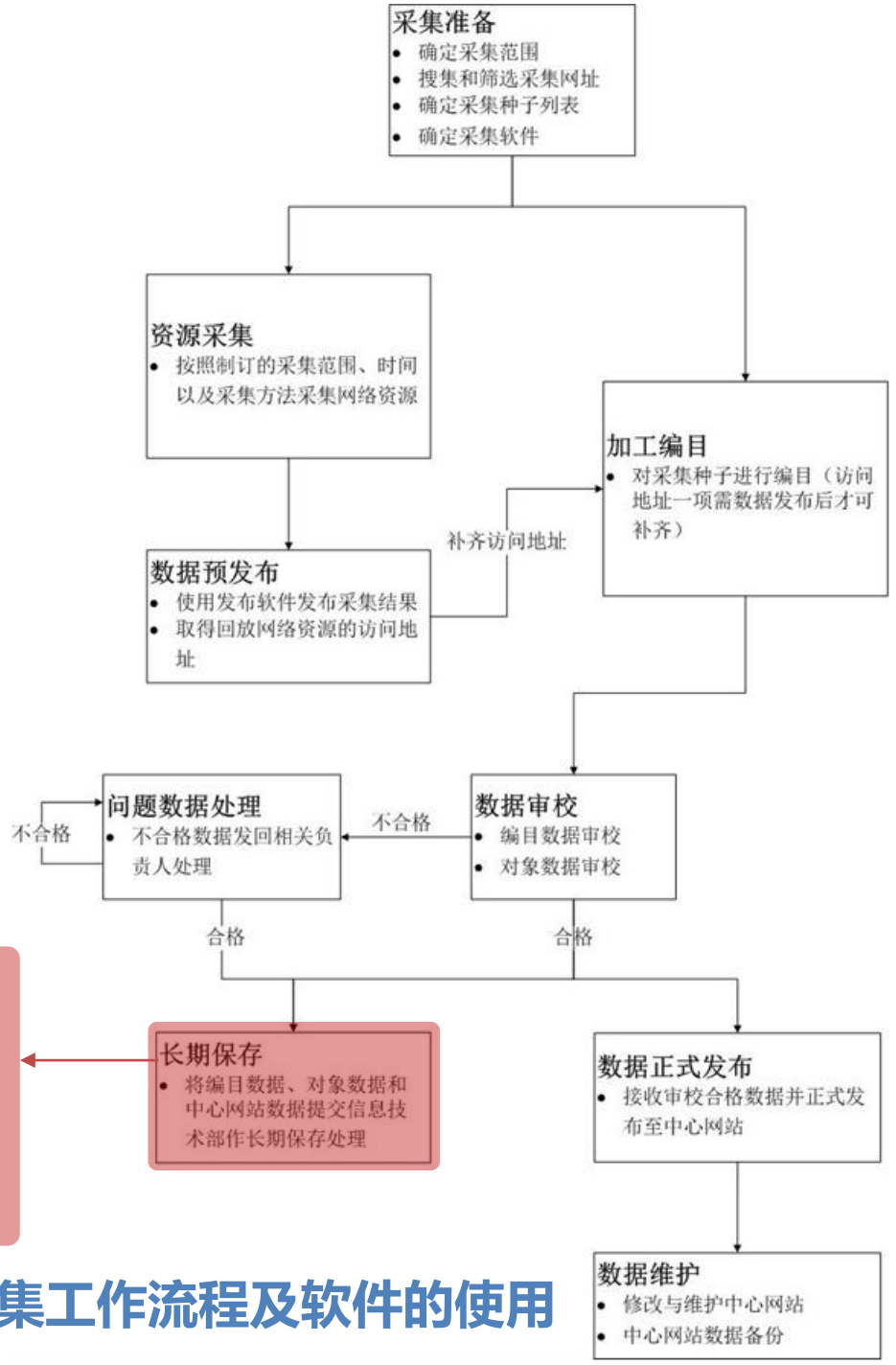
接收审校合格数据，将元数据导入数据库，通过中心网站进行正式发布，并提供用户服务。

（四）开源系统的网络资源采集工作流程及软件的使用



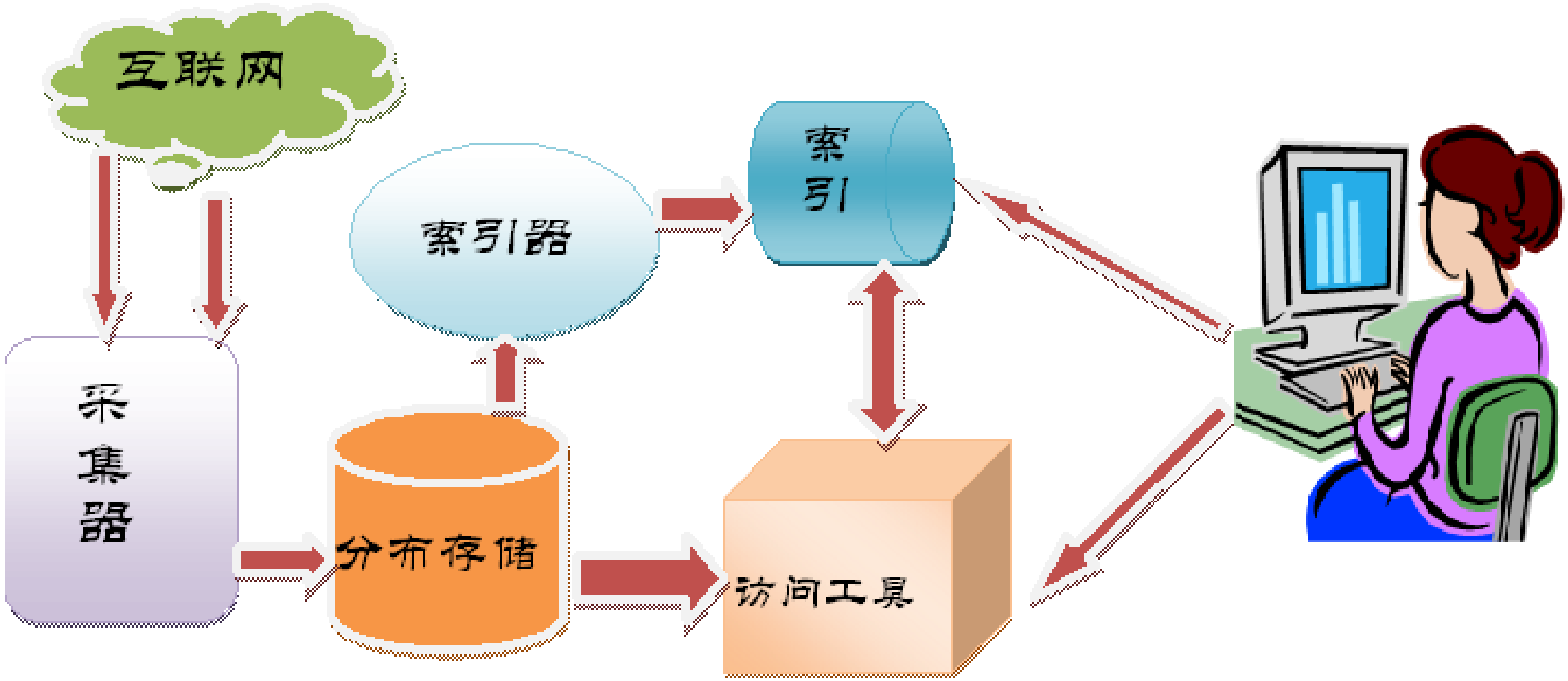
对中心网站页面进行修改与维护，保障数据准确无误，显示正常，同时做好所有数据的备份工作

（四）开源系统的网络资源采集工作流程及软件的使用



对所有成品编目数据、对象数据和中心网站相关数据按照信网部数据长期保存的要求进行压缩等处理，提交信网部进行长期保存

（四）开源系统的网络资源采集工作流程及软件的使用



（五）国家图书馆开发的网页资源采集与获取系统工作流程及软件的使用

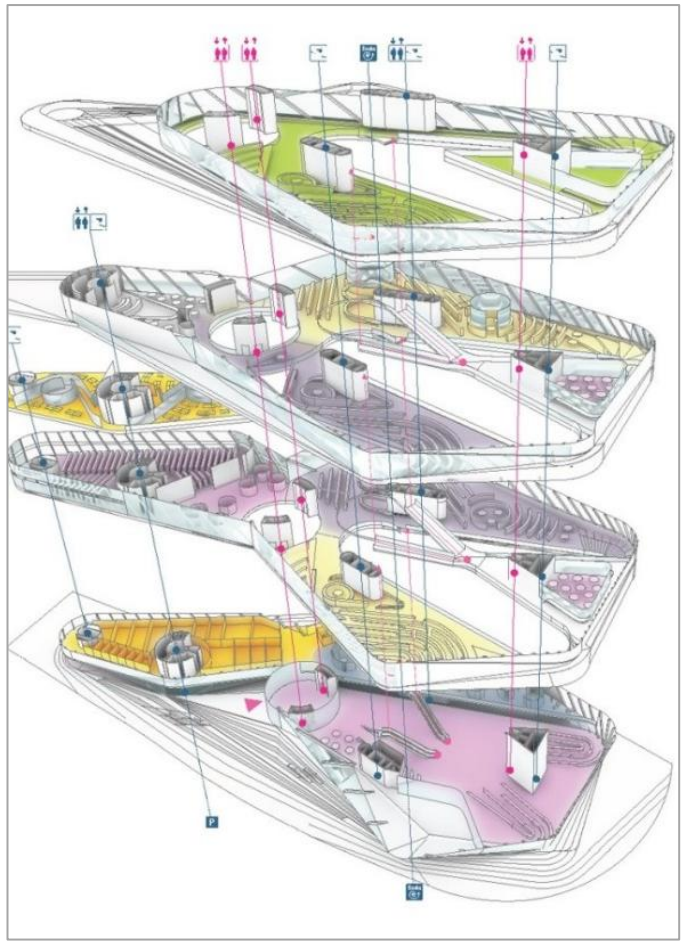
网页资源获取系统框架
网页资源获取系统部署在
CentOS 5.9 操作系统中

（五）国家图书馆开发的网页资

6

融合大数据理念
提炼网络信息资源重要价值

（一）加强顶层设计,建设全生命周期的网络信息保存保护体系



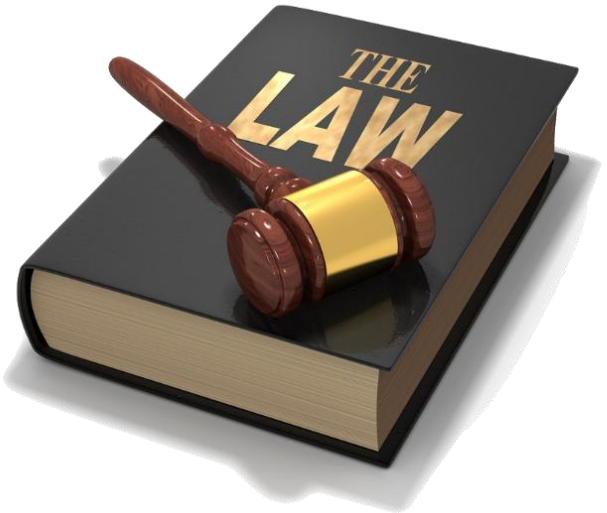
- 做好顶层设计，围绕网络信息生命周期构建完整的网络信息采集、保存、管理、分析、应用和服务模型并启动建设。以社会和用户需求为依据，对重要政治、新闻、文化、经济、科学、教育、安全领域的网络信息做重点采集和保存。

- 探索网络信息的分析、挖掘与组织技术，建设网络信息专题库。提高资源的服务水平，实现网络信息资源的一站式检索、科学化导航、可视化展示等。

- 要联合相关机构，共建共享网络信息保存保护工作成果，加强工作合力，提高整合效益。

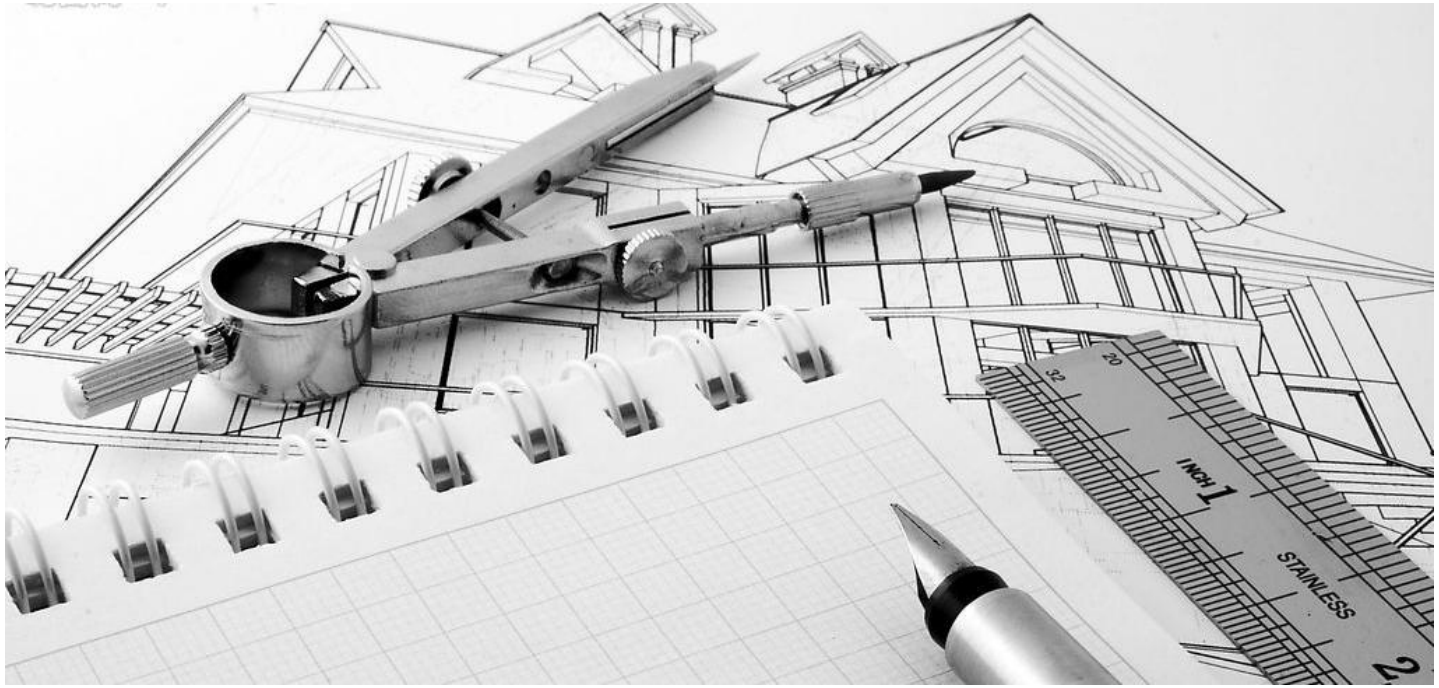
（二）推进法律制定，建设特色化网络信息保存法律体系

- 加强网络信息基础设施、网络信息技术发展、网络信息资源、网络信息安全等方面的政策建设
- 对《出版管理条例》等系列呈缴相关规定进行修订，围绕缴送范围、缴送方式、缴送期限、缴送格式、利用条件等问题进行详细的制度设计
- 推动对《著作权法》《信息网络传播权保护条例》等版权法规的修订，允许具有保存国家文化遗产职责的法定保存机构采集互联网上向公众开放且无获取限制的网络资源并进行长期保存。



（三）制定统一标准，提高工作效率和规范化管理

- 制定网络信息的采选和评价规范
- 制定网络信息元数据编目规则
- 制定严格的管理制度



（四）制定完整网络资源采集方案，形成有效、全面、特色的网络资源内容体系

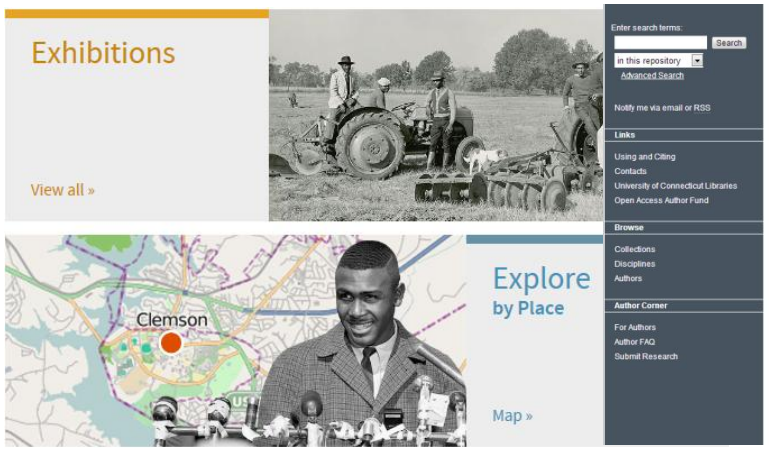
- 1、根据我国的政策规划，对国内重要领域网络资源做重点采集
  - （1）具有中华人文特色的和文化意义的网络资源
  - （2）与重大国家决策、立法、安全相关的网络资源
  - （3）与国家重大项目与工程建设相关的网络资源
  - （4）具有重要学术科研价值的网络资源
  - （5）国家重要政治、经济领域的网络资源
- 2、选择对我国有重要参考和安全价值的国外网站进行采集
  - （1）国外重要国家的主要行政部门的网站
  - （2）国外重要社交领域网站的采集
  - （3）国外经济领域重要项目、工程的网络资源
  - （4）国外先进科学技术、不同国家和地区文化的网络资源
  - （5）国外重要图情组织的网络资源，如IFLA、ALA等
- 3、根据网络资源的不同特征制定灵活多样的网络资源的采集方案

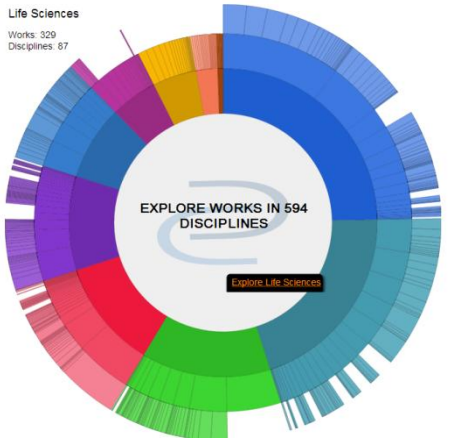
（五）重点研究并突破网络资源的采集、分析、管理与组织等关键技术

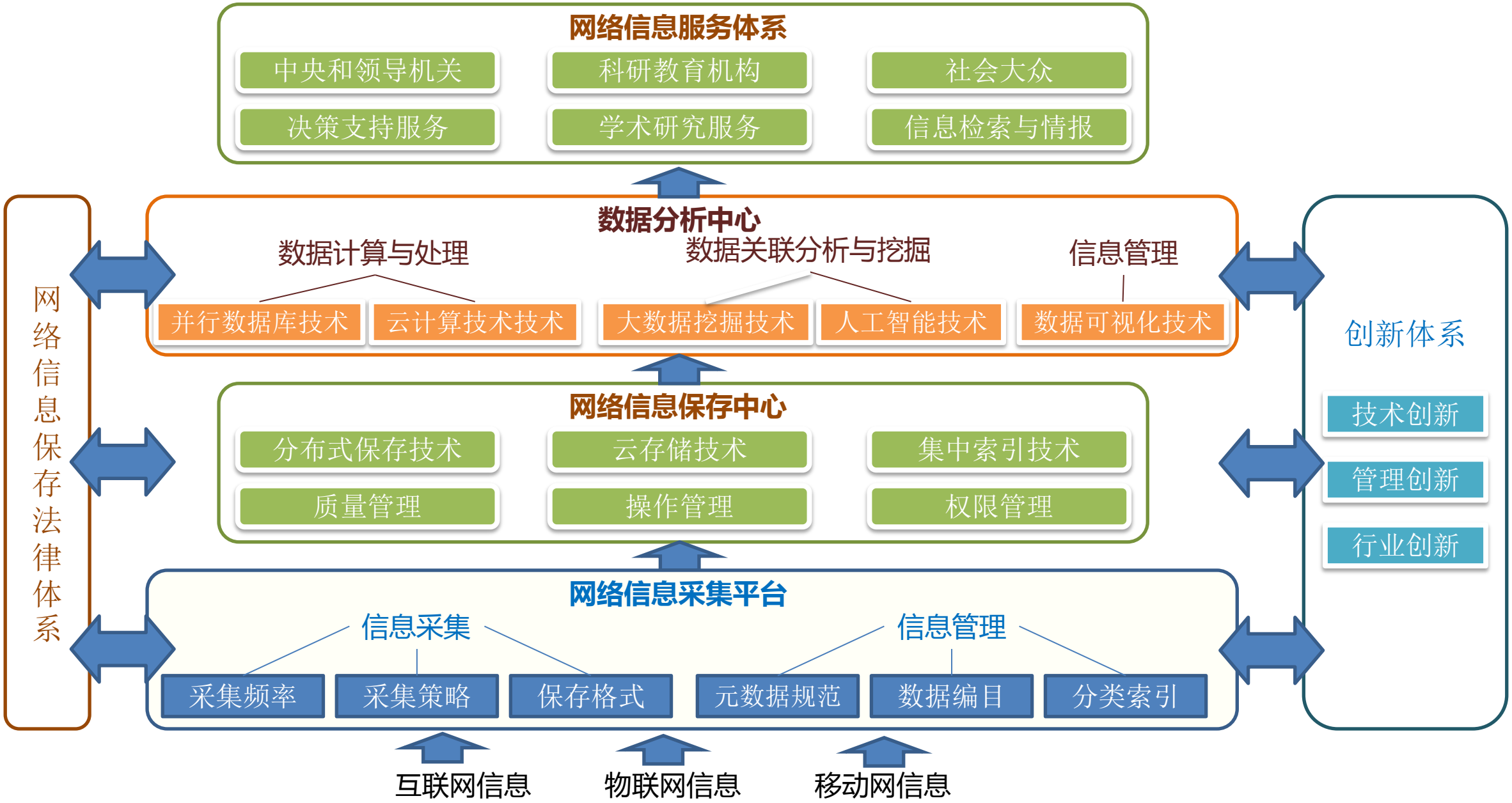
- 1、网络资源采集技术
- 2、网络存档资源的组织技术
- 3、网络存档的存取技术
- 4、网络存档资源的长期保存技术
- 5、网络存档资源的服务技术

（六）提高网络资源的发布与利用，满足用户多样化需求

- 1、建立国内外网站镜像，提供原生数据服务
- 2、建立多领域专题库，实现知识化服务
- 3、建立全文检索系统，实现一站式检索
- 4、通过可视化和智能化技术，提高资源揭示与展示水平






网络信息服务体系
中央和领导机关
科研教育机构
社会大众
决策支持服务
学术研究服务
信息检索与情报
数据分析中心
数据计算与处理
数据关联分析与挖掘
信息管理
并行数据库技术
云计算技术技术
大数据挖掘技术
人工智能技术
数据可视化技术
网络信息保存中心
分布式保存技术
云存储技术
集中索引技术
质量管理
操作管理
权限管理
网络信息采集平台
信息采集
信息管理
采集频率
采集策略
保存格式
元数据规范
数据编目
分类索引
网络信息保存法律体系
创新体系
技术创新
管理创新
行业创新
互联网信息
物联网信息
移动网信息

（一）建设一体化网络信息采集平台

（二）建设完善的网络信息保存中心

（三）建设高效的数据分析中心

（四）建设特色化网络信息保存法律体系

（五）带动技术与应用领域创新体系建设

（六）建设全国多层级网络信息服务体系

谢 谢！
Thanks