

# Web Information Preservation at National Library of China

中国国家图书馆的网络信息资源保存试验项目

Wang Zhigeng/王志庚  
National Library of China  
Email : wangzhg@nlc.gov.cn  
2004.07.15

# Outline

## 要点

---

- Brief introduction to NLC
- 国图简介
- WICP (**W**eb **I**nformation **C**ollection and **P**reservation) Project
- WICP试验项目
- Other efforts
- 其他



# Brief introduction to NLC

## 国家图书馆概况

---

- Established in 1909, a history of 95 years
- 始建于1909年，拥有85年的历史
- Open to the public on August 27, 1912
- 1912年8月27日对公众开放
- Started accepting legal deposit copies in 1916
- 1916年开始接受国内出版物的呈缴本
- A Collection of 24 million volumes
- 馆藏2411万册件，是亚洲最大的图书馆（2004.06）



Main building/总馆



Branch library/分馆



New building/新馆

# Key functions of NLC

## 国家图书馆的职能

---

- A deposit library
- 国家总书库
- The National Bibliography
- 书目数据中心
- References, reading and loan services
- 读者服务
- To see more at <http://www.nlc.gov.cn>



# Preservation of printed documents

## 纸质文献资料的保存和保护

---

- Adequate storage conditions
- 国际一流的善本书库和书刊保存本库房
- Microfilming and digitizing
- 馆藏文献的缩微化和数字化



# Preservation of digital information

## 数字信息长期保存和保护

---

- Not easy, different from the preservation of printed materials
- 与印刷型文献的保存不同
- Legal deposit policies, intellectual property rights
- 缴送制度、著作权等法律和制度问题
- Preservation of digital environment and metadata
- 数字对象的保存、数字信息环境的保存、保存元数据等技术问题
- Organizational, social and economic issues
- 组织机构、社会分工和经济模型问题



# Our Understanding in Web Archiving

## 国家图书馆的基本认识

---

- Web information resources have been a major part of Chinese civilization and digital heritage, which should be properly preserved and protected.
- 网络信息资源是中华文明成果，是中华数字文化遗产的一部分，应该得到妥善地保存和保护。
- They bear a strategic meaning for NLC's collection development and public services. The NLC should collect web information resources as what has been done for paper-based materials.
- 网络信息资源对国家图书馆的馆藏发展和服务具有战略意义，国家图书馆应该像收集传统的图书资料那样，全面收集各种网络信息资源。





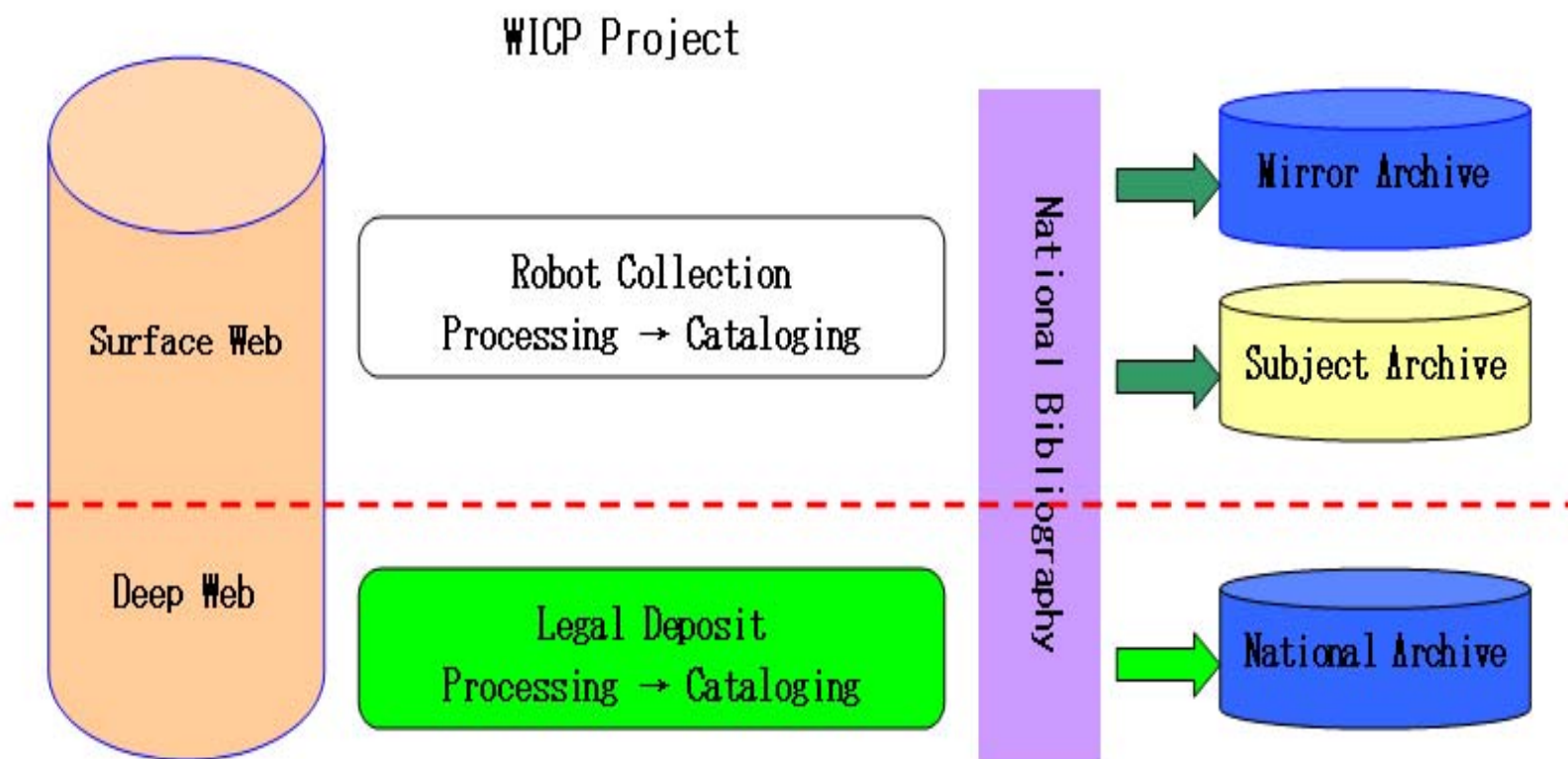
# Why preserve web information? 为什么保存网络信息？

---

- Increasing masses of information published through the web  
越来越多的信息以web的形式发布
- Volatility of web information, the average life span of web pages is 75 days.  
Web的挥发性，网页的平均寿命为75天
- A new dimensional space for social culture  
社会文化的一个新维度空间，是现代的文化遗产
- Many early web pages have disappeared  
早期的web信息已经消失！

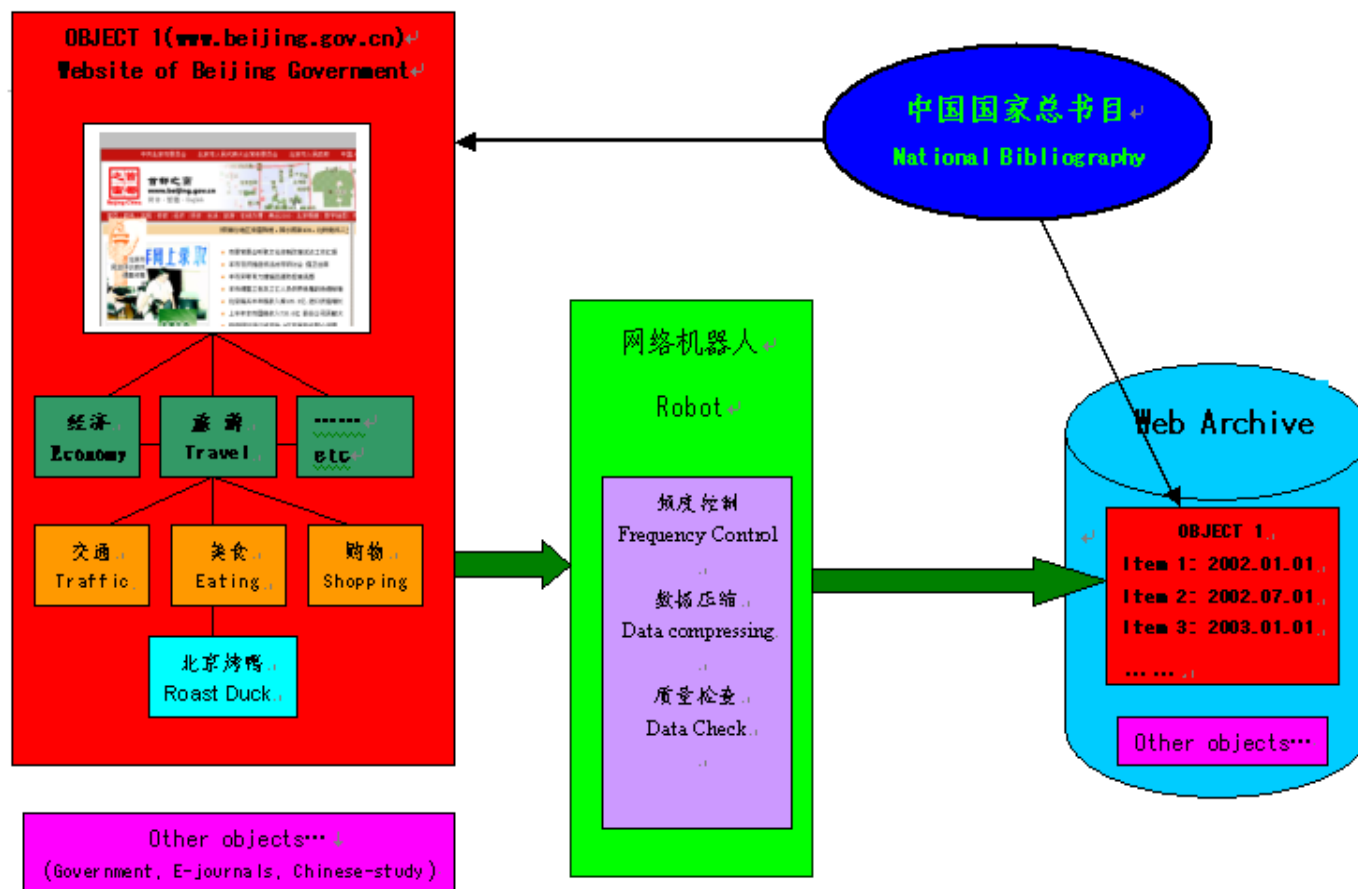
# WICP model

## WICP 示意图



# Mirror Archive

## 镜像存档





# Mirror Archive

## 镜像存档

---

### ■ Work flow/业务流程

1. Survey of target website/目标网站调查
2. Capturing conditions/采集条件设定
3. Starting capture/开始采集
4. Cataloging/元数据编目
5. Quality control/质量检查
6. Downloaded website registry/信息单元登记
7. Providing service/提供服务



# Mirror Archive

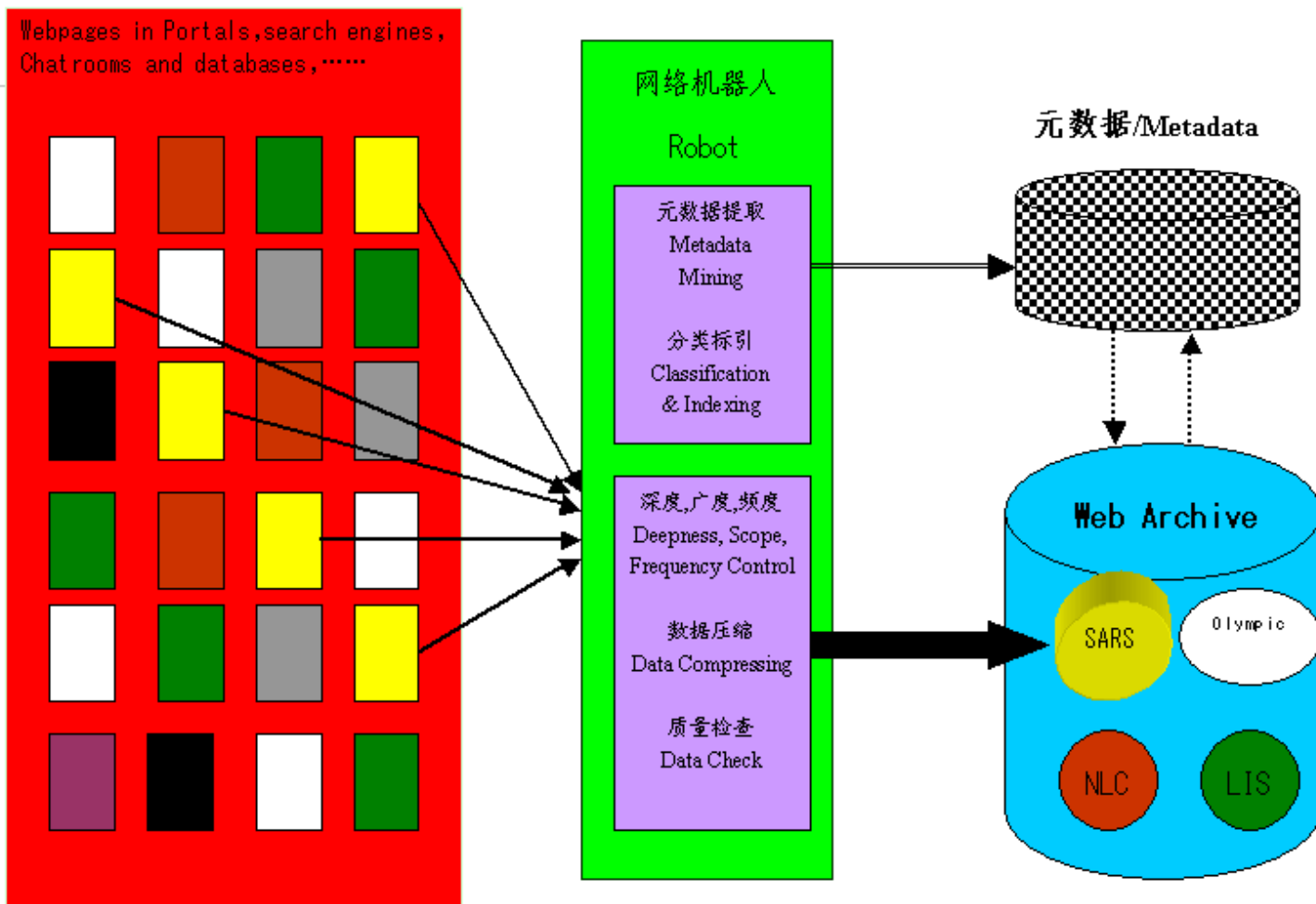
## 镜像存档

---

- **Collection statistics/馆藏情况**
  - **Government Information (.gov)/政府网站**
  - **E-journal /电子报刊**
  - **Chinese Studies/中国学**

# Subject Archive

## 专题存档





# Subject Archive

## 专题存档

---

- Workflow/业务流程
  1. Selection of subject/主题的选择
  2. Survey/对象调查
  3. Capturing conditions/设定采集条件
  4. Starting capture/开始采集
  5. Metadata mining/元数据挖掘
  6. Object downloading/网页快照
  7. Data storage/数据保存
  8. Quality control/质量检查
  9. Providing service/提供服务



# Subject Archive

## 专题存档

---

- **Collection statistics/馆藏情况**
  - **2008 Beijing Olympic Games/ 2008北京奥运会**(ongoing)
  - **SARS/非典专题**(Finished)
  - **The manned space flight project/中国载人航天工程**  
(Finished)
  - **Media report about NLC/国家图书馆媒体报道**(ongoing)
  - **Library studies and information science/图书馆情报学**  
(ongoing)





# Some issues to be addressed

## 存在的问题

---

- Web robot/网络机器人技术
- Storage/海量信息存储技术
- .....



## Other efforts

### 其他工作

---

- **Suggestions for policy-making**
- 建议网络信息资源作为呈缴对象写进《中国图书馆法》
- **Technological attempts of digital information preservation, such as reformatting and migration**
- 数字信息资源保存技术方面的尝试



# Cooperation

## 合作

---

- Although there has not come up with an effective way to preserve digital resources, we are ready to work with all colleagues in library community to preserve digital information.
- 目前还没有一种行之有效的数字资源的保存技术和策略，我们愿意和国内外的所有同仁一起为保护人类共同的数字文化遗产而努力。



---

**Thank you !**  
**谢 谢 !**