Developing a Highly Automated Web Archiving System Based on IIPC Open Source Software

Zhenxin Wu, Jing Xie

Jiying Hu, Zhixiong Zhang

National Science Library, Chinese Academy of Sciences

iPres2015,Chapel Hill,November

outline

- 1 Introduction
- 2 Developing Web Archiving System
  - Web Archive Needs of NSL
  - Web Archiving System Framework
  - System Function Framework
  - Automated Workflow
- 3 Current Progress
- 4 Next Developing Plan

1.Introduction

- **Preserving online science information has explicitly become a national strategy.**

------National Digital Information Infrastructure and Preservation Program. 2012. Science @ Risk: Toward a National Strategy for Preserving Online Science. Library of Congress, Washington, DC.

1.Introduction

- **National Science Library (NSL), Chinese Academy of Sciences (CAS)**
  - began a two-years pilot project with supporting by National Social Science foundation of China in 2006
  - Got another two-years funding from CAS to develop an operating system (**NSL-WebArchive**) for archiving the important web information in 2013.

Outline

- 1 Introduction
- 2 Developing Web Archiving System
  - Web Archive Needs of NSL
  - Web Archiving System Framework
  - System Function Framework
  - Automated Workflow
- 3 Current Progress
- 4 Next Developing Plan

2.1 Web Archive Needs of NSL

- Harvest periodically and sustainably
- Balance harvest frequency and speed so that it will not affect daily access of seed sites.
- Want more metadata and management
- Highly automated workflow to reduce manual work
- Support in-depth analysis of archived data
- Provide more services for users based on archived data

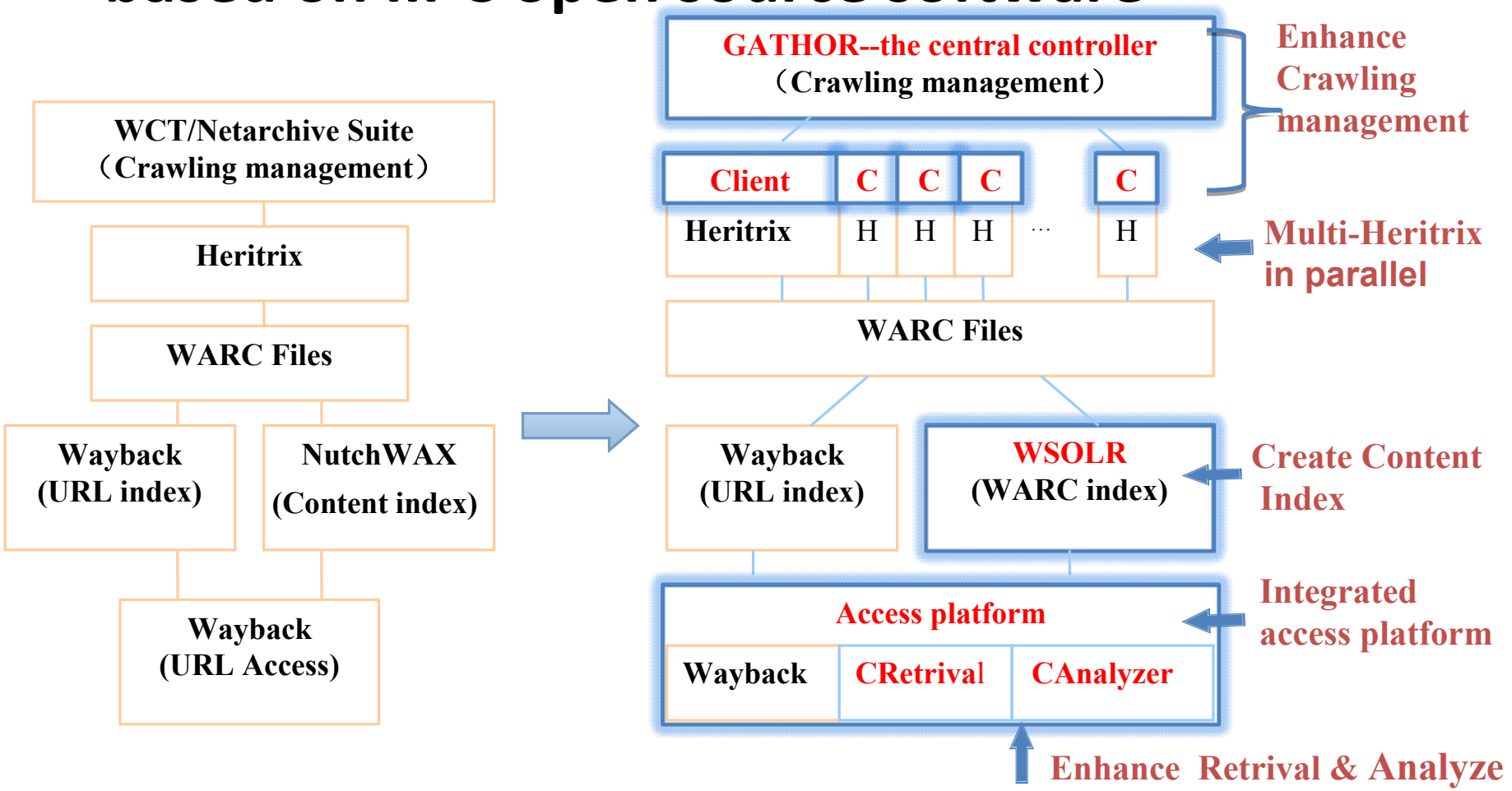
A high-performance system with less developing invest

2.2 Design Web Archiving Framework

- **before developing web archiving system**
  - **An investigation of IIPC web archiving tools (The International Internet Preservation Consortium)**
    - **Heritrix**, a highly-scalable crawler created by the Internet Archive.
    - **Web Curator Tool & Netarchive Suite** , Crawling management tool.
    - **Wayback**, an index and access tool based on URL.
    - **NutchWAX**, a full-text index tool.
  - **Some research on other libraries' work**
    - **French National Library**
    - **British Library**
    - **National Library of China**
    - **……**

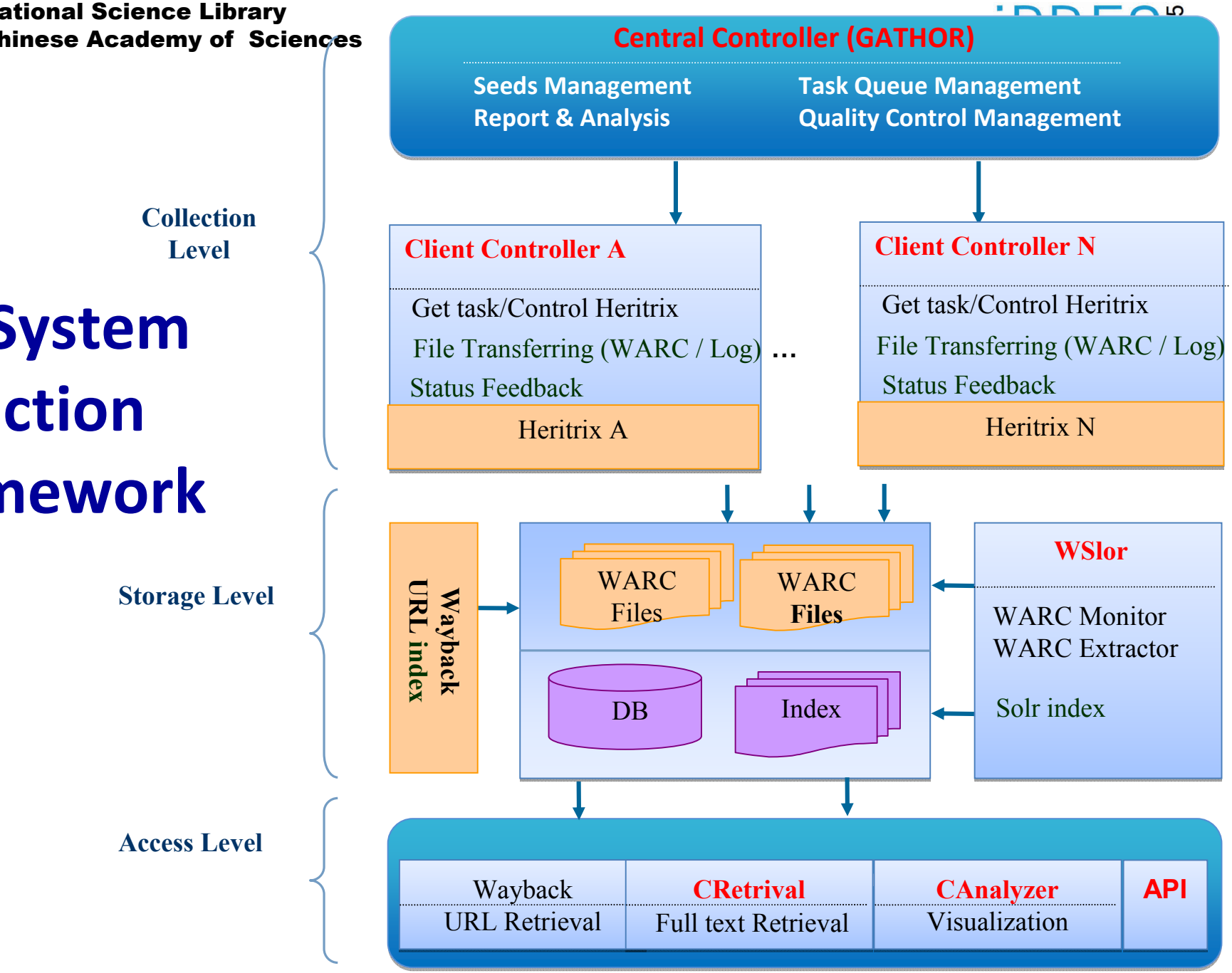
2.2 Design Web Archiving Framework

- based on IIPC open source software



IIPC Web Archiving Framework NSL Extension Framework Base on IIPC's

2.3 System Function Framework



2.3 System Function Framework

- Collection Level (**Central Controller-GATHOR**)
  - **Seeds Management**: more described metadata(including type, subject, domain), administrative metadata and configure info.
  - **Crawling Task Queue Management**: automatically generate and schedule the crawling task, and monitor the status of each task.
  - **Report & Analysis** : analyze crawling log and provide analysis report
  - **QC Management**: check error for improving crawling effect

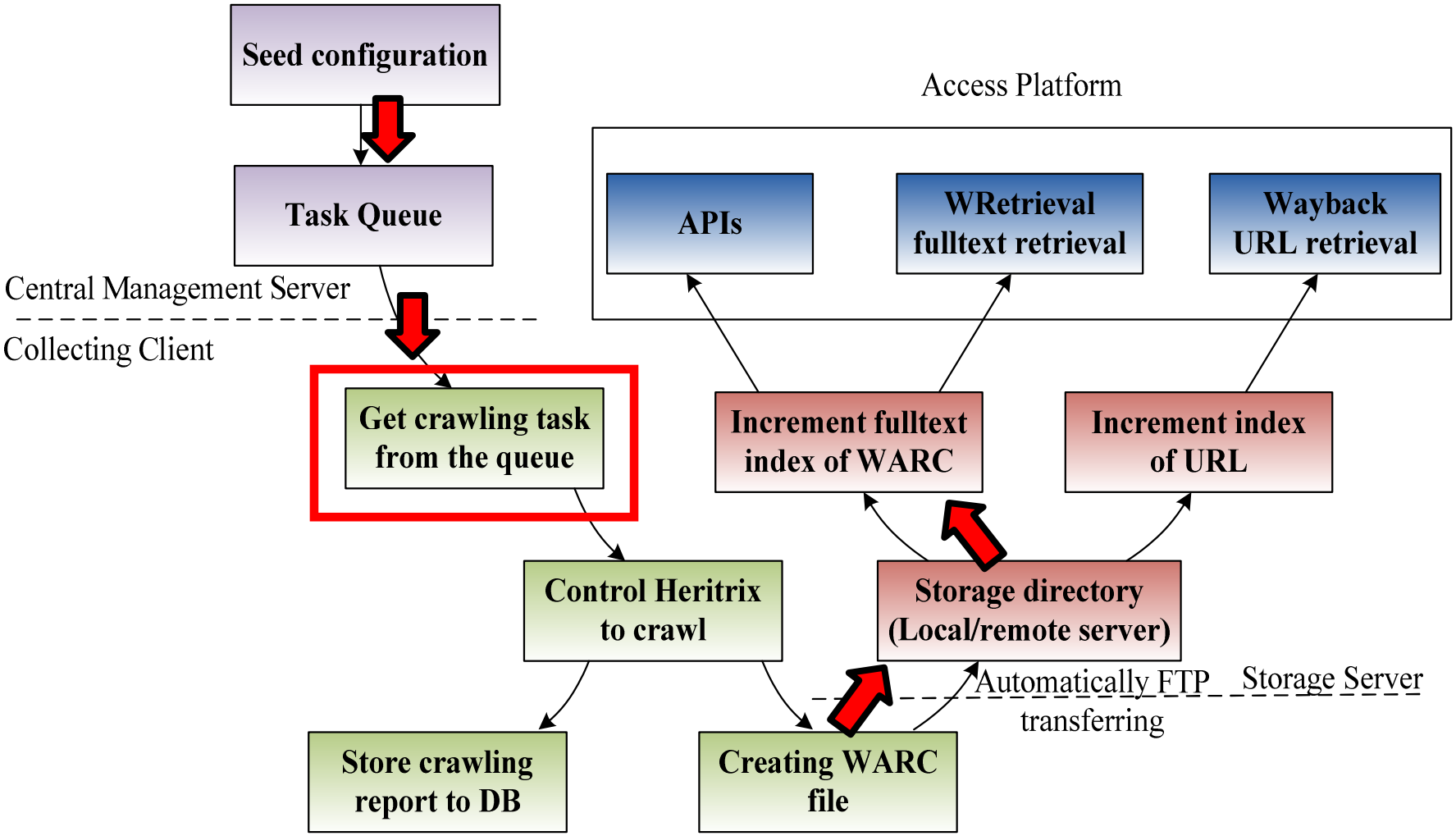
2.3 System Function Framework

- Collection Level (**the Collecting Client**)
  - **Task Control Module:** gets a task ,controls Heritrix to crawl web and monitor it's status.
  - **File Transferring Module** : transfer WARC files and crawling logs to the specified directory in the remote storage server
  - **Status Report Module**: Report task status to GATHOR

2.3 System Function Framework

- Storage level
  - **WSOLR** with three sub-modules
    - **WARC Monitor**: automatically monitor the specified directory for the new uploaded file
    - **WARC Extractor**: extract related information from these files
    - **SOLR** : create incremental Solr index
- Access level
  - **CRetrival** : provide full text retrieval and facet navigation.
  - **CAnalyzer** : provide statistic and analysis function

2.4 Automated Workflow of NSL-WebArchive



2.4 Automated Workflow of NSL-WebArchive

Three key parts for highly automated workflow

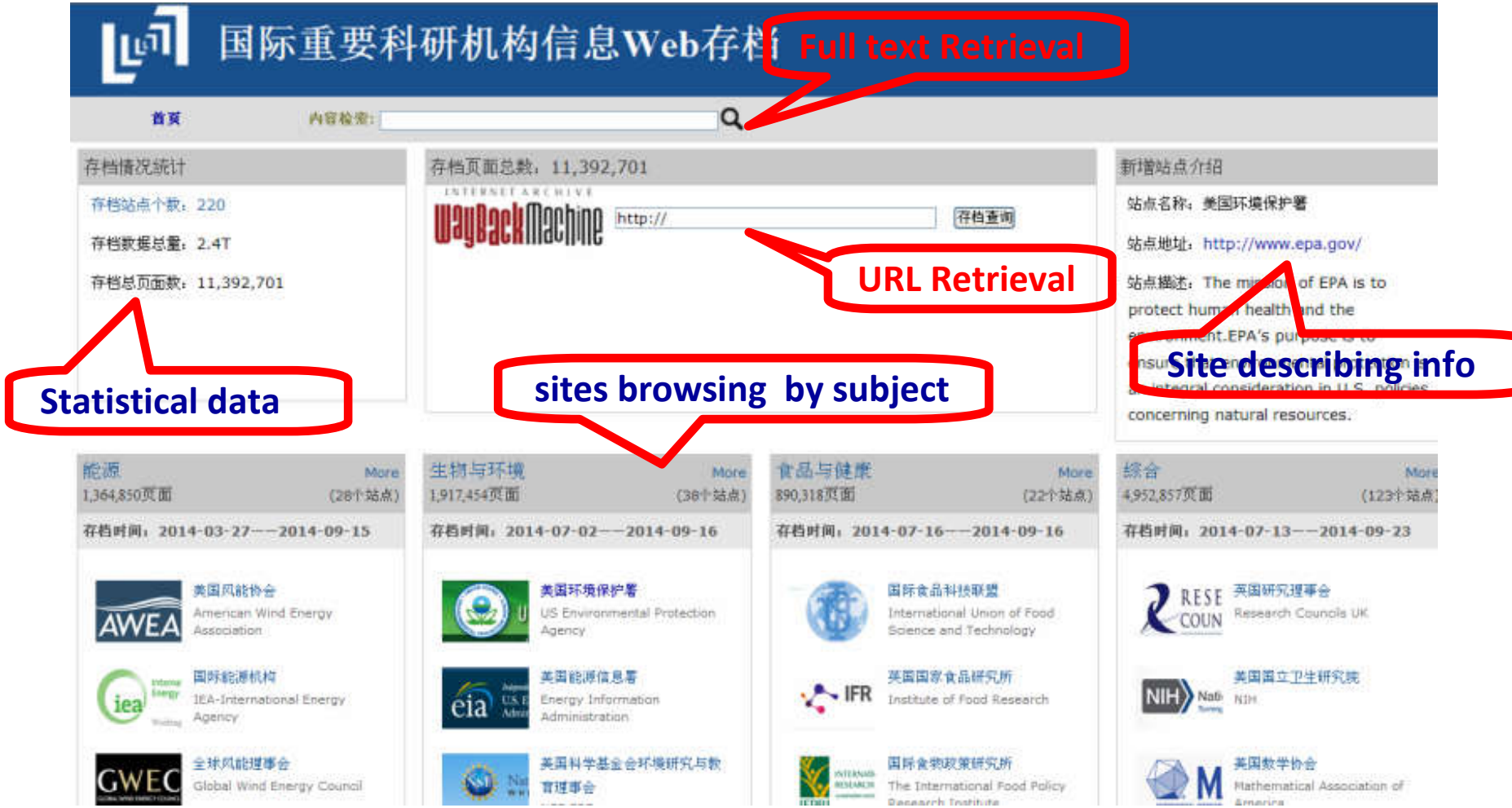
- **GATHOR**
  - **Task Queue Management Module** : automatically generate and schedule the crawling task
- **Collecting Client**
  - **Task control module, File transferring Module, Status Report Module**: automatically get a task ,control Heritrix and transfer files.
- **WSOLR**
  - **WARC monitor, WARC extractor, SOLR**: automatically monitor WARC files , extract data and create index

3. Current progress

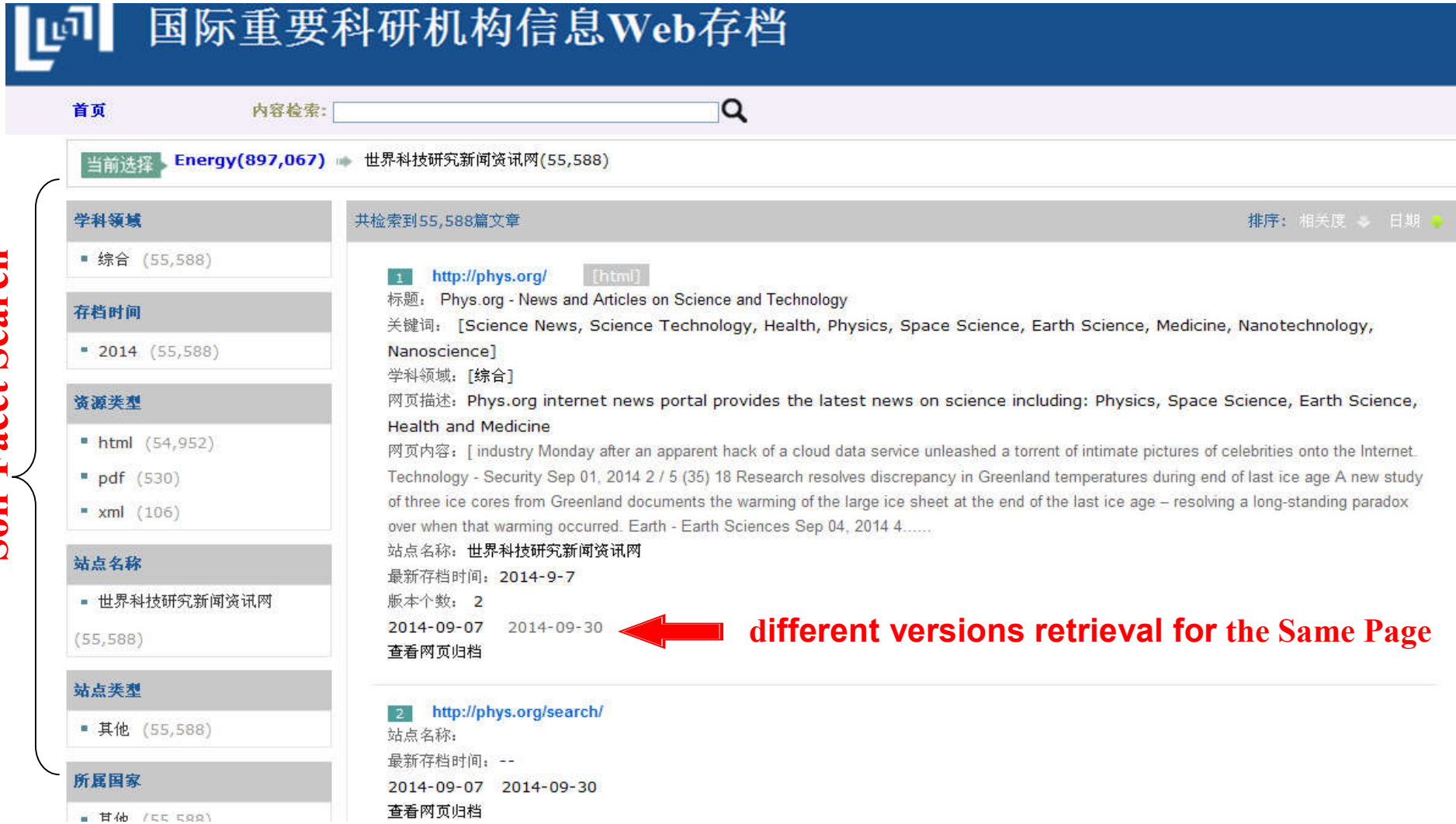
Have finished the first stage work of system developing

- Enough metadata and more effective management
- A highly automated web archiving workflow
- More access services for users
  - URL retrieval, full text retrieval and facet navigation, Some statistical functions, Sites browsing

An integrated access platform



Full text retrieval and facet navigation



Some statistical analysis

站点名称：欧盟生物燃料技术平台

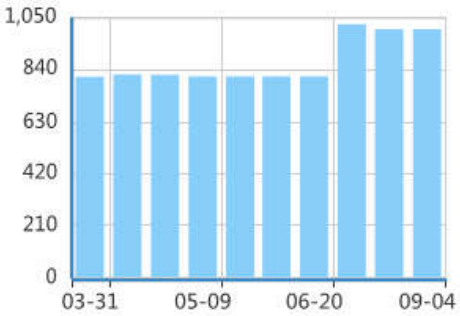
站点地址：http://www.biofuelstp.eu/

学科领域：能源

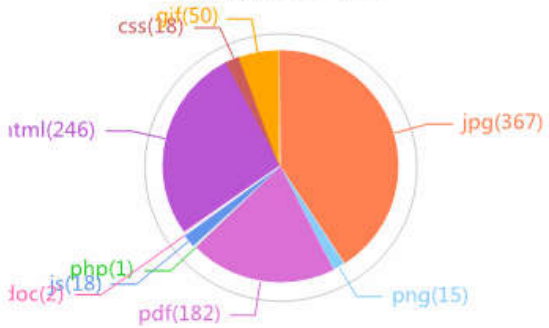
存档记录：

序号	存档时间	URL数	存档数据量	查看
1	2014-03-31	809	150370247 (143 MB)	查看
2	2014-04-10	818	150461460 (143 MB)	查看
3	2014-04-25	816	150379322 (143 MB)	查看
4	2014-05-09	810	148856813 (142 MB)	查看
5	2014-05-10	810	148856962 (142 MB)	查看
6	2014-06-04	810	148856962 (142 MB)	查看
7	2014-06-20	811	148850413 (142 MB)	查看
8	2014-07-19	1022	181303335 (173 MB)	查看
9	2014-08-04	1006	181726782 (173 MB)	查看
10	2014-09-04	1006	181734512 (173 MB)	查看

存档时间-URL数分布图



存档种类分布图



4. Next developing plan

- How to gain maximum value from archived resources
  - Support in-depth data mining
  - Provide functions for effective assessment of S&T policy and technology decisions, strategic decisions, trends analysis of domain, and predict future trends, etc.

These needs will become the main target of our next developing plan

Thanks

Zhenxin Wu

wuzx@mail.las.ac.cn

Jing Xie

xiej@mail.las.ac.cn