大数据众包：CADAL联盟共享模式探讨

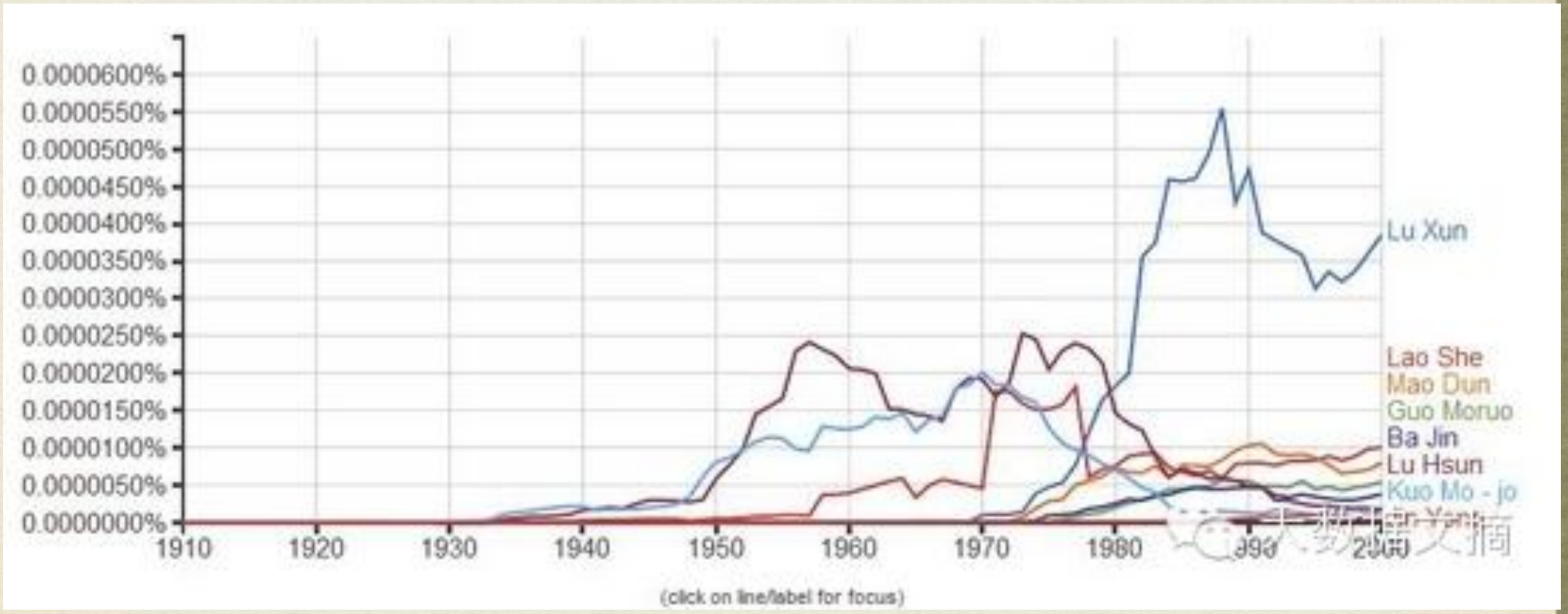
CADAL项目管理中心·黄晨

2014.7 @ 长春

各种大数据

《计算历史学：大数据时代的读书》

——尼克，东方早报



各种大数据

- 全世界性服务工作者：4200万，如果由TA们组成一个国家，将成为全球人口排名第31位的国家。
- 高达1/8的网站都是色情网站。ManWin，控股全球排名第一的成人站，每个月能拿到大约16亿浏览者的数据资料
- 每秒钟全球有28000人在同时浏览色情网站。
- 搜索引擎中高达1/4的搜索请求与色情相关。
- 35%的网络下载是色情内容。

各国色情内容消费额排名

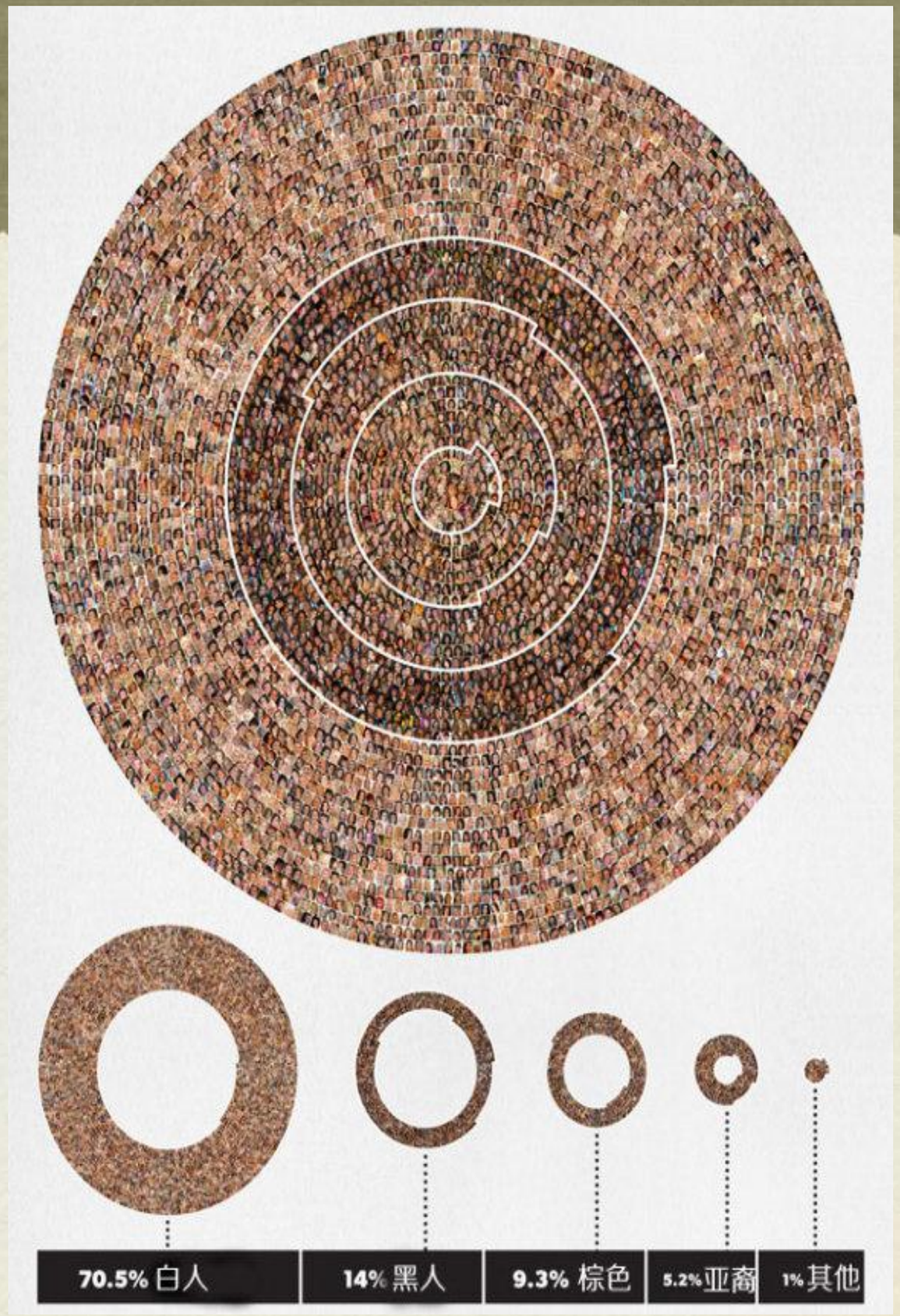
人均色情内容消费额排名

各种大数据

Jon Millward

6个月的时间，潜心分析超过10000名的色情艳星及她们的12万部作品。

- 头发以棕色的最多：**39.1%**
- 其次金色：**32.7%**
- 黑色：**22.5%**
- 红色：**5.3%**



各种大数据

HUSTLER MAGAZINE · HUSTLER HOLLYWOOD · HUSTLER CLUB · HUSTLER TV · HUSTLER CASINO · HUSTLER toys · HUSTLER.CLOTHING

工作人员在TW上无意发现很多人居然在问男主角的衣服是什么牌子的，在哪里可以买到。后来他们就引进了服装生产线，开始生产Hustler品牌的个性时尚服饰。原本他们担心自己成人制造商的口碑会影响销路，但是数据监控发现他们位于好莱坞和纽约的实体店，60%的顾客是女性，其中很多人并不知道他们之前是干什么的，并且有不少好莱坞明星会主动穿着他们的衣服亮相。

http://hustlerclothing.com/wp/

各种大数据

- 柜台和店内各角落都装有摄影机，
- 店经理随身带着PDA。
- 销售人员结帐、盘点每天货品上下架情况，并对客人购买与退货率做出统计。
- 结合柜台现金资料，交易系统做出当日成交分析报告，分析当日产品热销排名，然后，数据直达Zara仓储系统 。

各种大数据

- 沃尔玛在全球超过200万名员工，总共有110个超大型配送中心，每天处理的资料量超过10亿笔。

- 2011年4月，沃尔玛以3亿美元高价收购了一家专长分类社群网站Kosmix。Kosmix不仅能收集、分析网络上的海量资料（大数据）给企业，还能将这些资讯个人化，提供采购建议给终端消费者。

各种大数据

- 亚马逊即将推
  —“需求方平
  Platform，D
  消费者相遇。
- 预测性物流专利：结合数据和指标预测顾客要下单的商品，减少顾客下单到收货之间的时间差，提高满意度。

数字图书馆的大数据启示

- 匹配　：了解读者需求

- 关联　：挖掘读者需求

- 创造　：预测读者需求

数字图书馆的大数据建设

- Internet Archive：6000TB，20TB/月
- SUMMON：800，000，000条
- CADAL项目：2，500，000册
- 中文期刊篇名：30，000，000条
- 中国学者SCI收录：2，100，000篇
- 中国专利：6，910，000件
- 作者数据：27，000，000
- 机构数据：2，750，000
- 在校大学生：26，480，000
- 今年毕业生：突破7，000，000

CADAL项目建设周期

- 项目一期建设100万册(件)数字资源，国家投入7000万元，美方合作单位投入约200万美金，“十五”期间已经完成。

- 二期建设在一期百万册的基础上，完成150万册(件)数字资源，在全国建设八个数字图书馆数据中心，实现数据安全和全球服务，由国家投入1.5亿建设资金，在“十一五”期间完成。

- CADAL项目从三期建设开始,在继续扩大资源建设的同时，在资源整合的基础上实现知识重构和信息创新，形成集资源采集、信息重组、内容创新、按需发布、个性服务为一体的学术数字图书馆体系。

CADAL项目建设回顾

- 2012年5月28日，CADAL项目二期通过了由教育部组织的专家组验收。
- 实现数字化资源250万册，保持公益性数字图书馆的领先地位
- 40余所高校建立数字资源加工中心，形成2个专业数字化加工基地。月加工能力：2100万页，7万册图书。
- 与Internet Archive在深圳保税区建立了国际合作扫描中心，是全世界最大的专业数字化加工基地。

CADAL项目建设回顾

项目参建单位增至70个

- 清华大学图书馆
- 北京大学图书馆
- 吉林大学图书馆
- 西安交通大学图书馆
- 浙江大学图书馆
- 复旦大学图书馆
- 南京大学图书馆
- 武汉大学图书馆
- 中山大学图书馆
- 四川大学图书馆
- 上海交通大学图书馆
- 北京师范大学图书馆
- 华中科技大学图书馆
- 中国科学院国家科学图书馆
- 中国人民大学图书馆
- 中国农业大学
- 中国科学院文献情报中心
- 新疆石河子大学
- 华南理工
- 暨南大学
- 香港大学
- 香港中文大学
- 香港城市大学
- 西南政法大学
- 华东师范大学
- 新疆农业大学
- 山东大学
- 汕头大学
- 北京交通大学
- 西南交通大学
- 电子科技大学
- 南昌大学
- 苏州大学
- 中央广电大学
- 北方民族大学
- 广西大学
- 重庆大学
- 哈尔滨工业大学
- 中央美术学院
- 中国美术学院
- 内蒙古大学
- 华中师范大学
- 西北工业大学
- 中国海洋大学
- 湖南大学
- ……

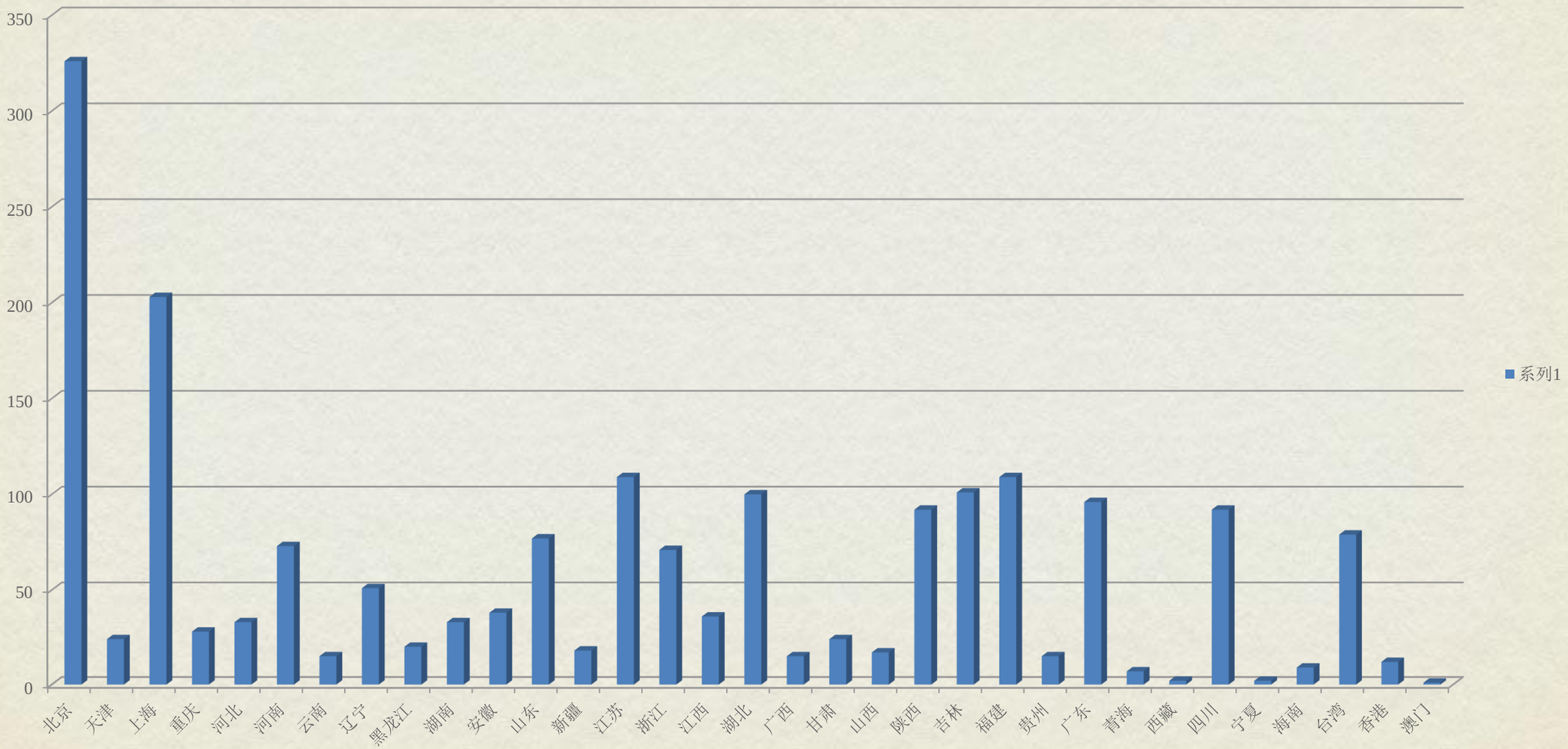
项目建设回顾

项目合作单位

- 卡内基-梅隆大学（CMU）
- 伊利诺斯大学香槟分校（UIUC）
- 哈佛燕京图书馆
- 哥伦比亚大学图书馆
- 斯坦福大学图书馆
- 普林斯顿大学图书馆
- 芝加哥大学图书馆
- 康奈尔大学图书馆
- 加拿大英属哥伦比亚大学图书馆
- 香港城市大学图书馆
- 香港大学图书馆
- 香港中文大学图书馆
- 埃及亚历山大图书馆
- 德国柏林国家图书馆
- Oxford University Press
- Internet Archive
- 印度科学院（班加罗尔）
- 教育研究网（德里）
- 安那大学
- 阿鲁密工程学院
- 果阿大学
- 印度天体物理学院
- 印度信息技术学院（阿拉哈巴德）
- 国际信息技术学院（海得拉巴）
- Kanchi Mutt
- 马哈拉施特拉邦工业发展合作组织
- 旁遮普技术大学
- Shanmugha科技艺术研究院
- 海得拉巴省城市中心图书馆
- 印度神祈巴拉吉的神殿
- 浦那大学
- ……

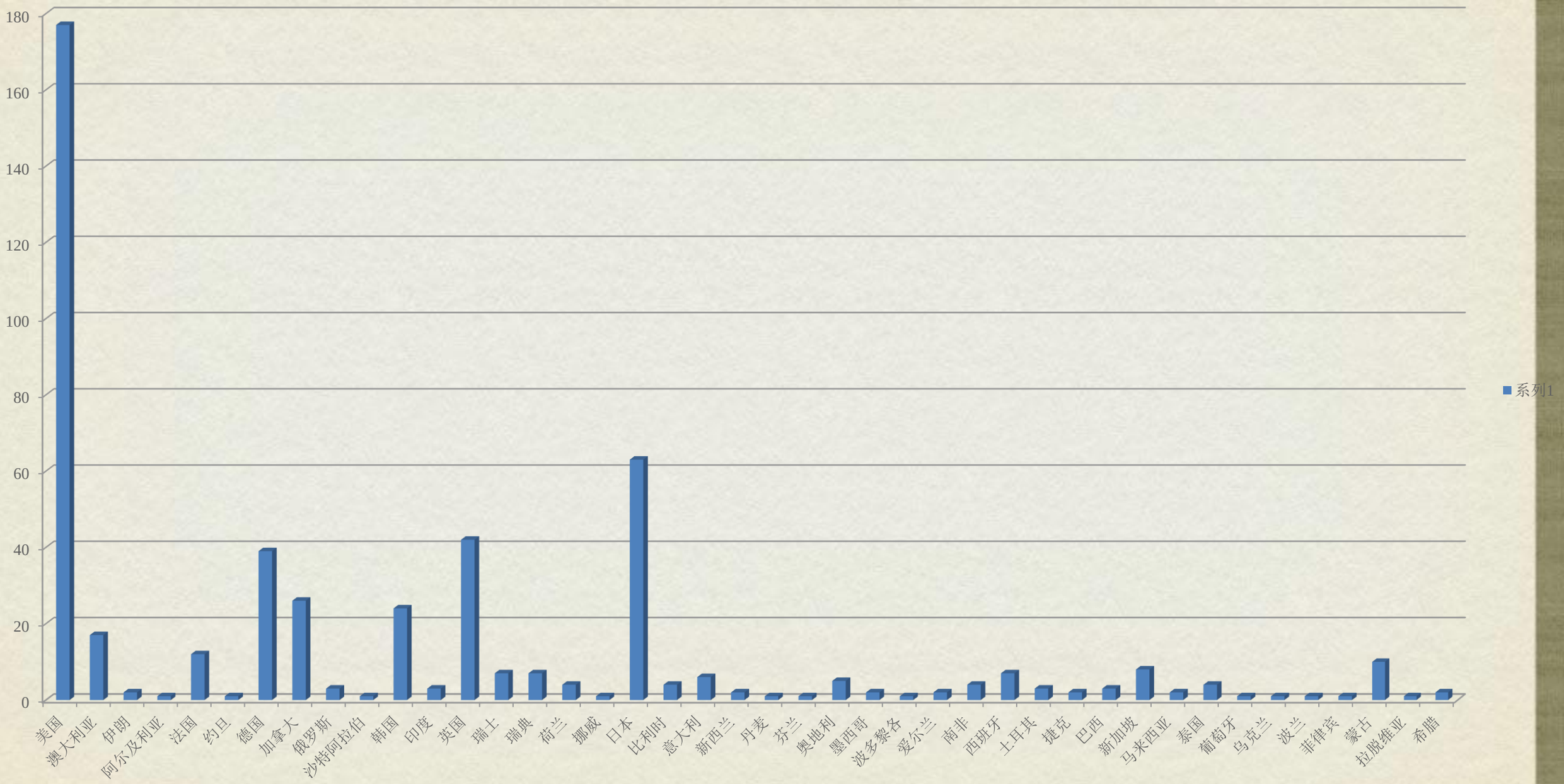
CADAL门户访问情况（2013年度）

- **国内1928个学校获取CADAL服务**



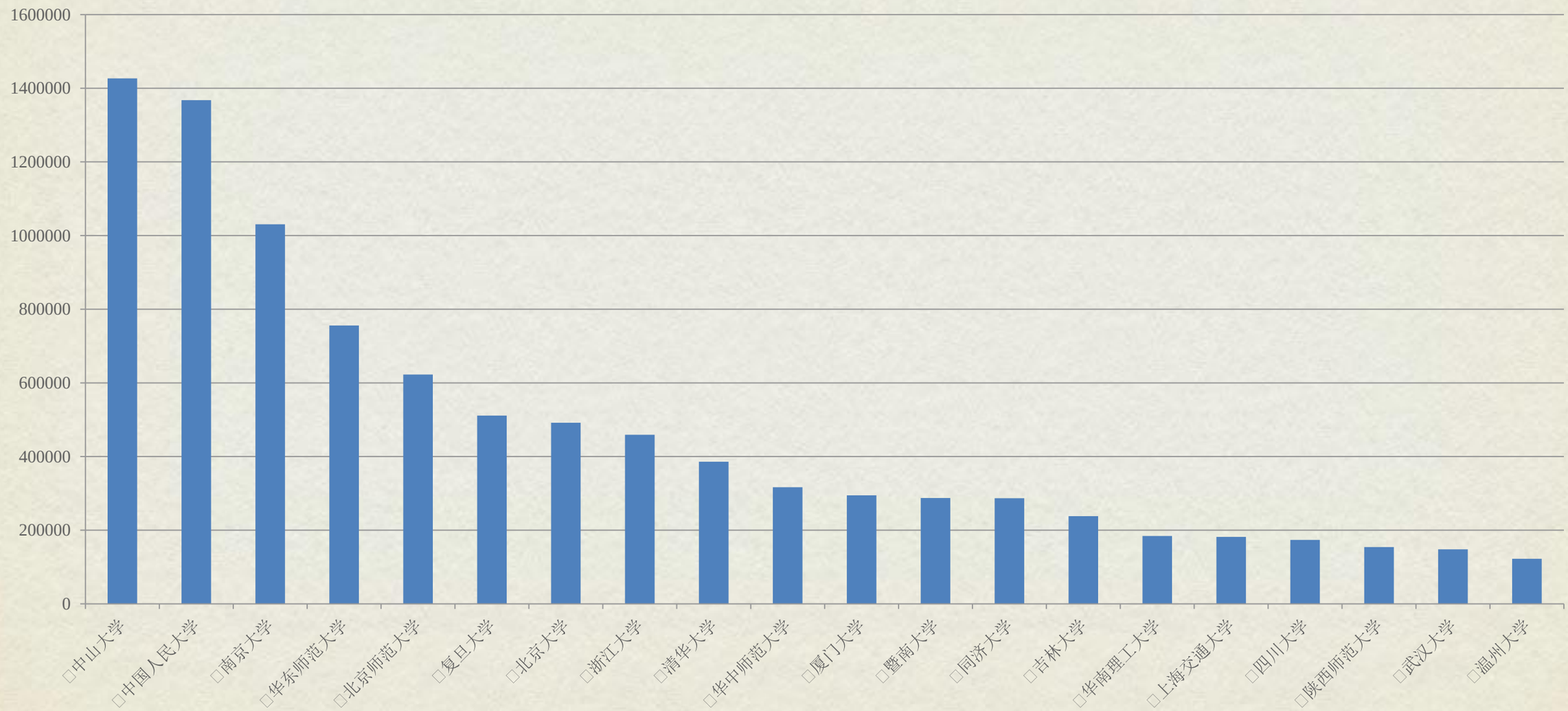
CADAL门户访问情况（2013年度）

国外504个学校访问CADAL门户



CADAL门户访问情况（2013年度）

- **阅读总量前二十的高校，共944万册（次）**



未来建设规划

- 项目建设思路
  - 通过大开放，构造大服务，利用大数据，实现大智慧。
  - 与用户交互共享知识，与信息机构透明共享用户。

- 项目建设内容
  - 整合海量资源
  - 融合先进技术
  - 泛在个性服务
  - 全球开放合作

建设规划：大数据下的资源建设

- 核心挑战之一是处理海量的非结构化数据。 这种新数据呈现与大海类似的性质，所以称之为数据海。

互联网
- 全球网站数量突破5亿
- 全球博客数量已达1.81亿
- 中国网站数量为230万，网页数量为866亿个

物联网
- 思科预测，到2020年，物联网中的物体将达到500亿部。

档案数据
- 至2011年底，全国共有档案馆4107个，已开放各类档案10376万卷（件）

图书期刊
- CADAL图书：250万册
- Google图书：1500万册
- 万方：期刊6000种，学位140万篇，会议论文180万篇

专利标准
- 至2011年底，共有274.0万件 专利
- 2011年制定和修订国家标准1993项，其中新制定1559项

科技报告

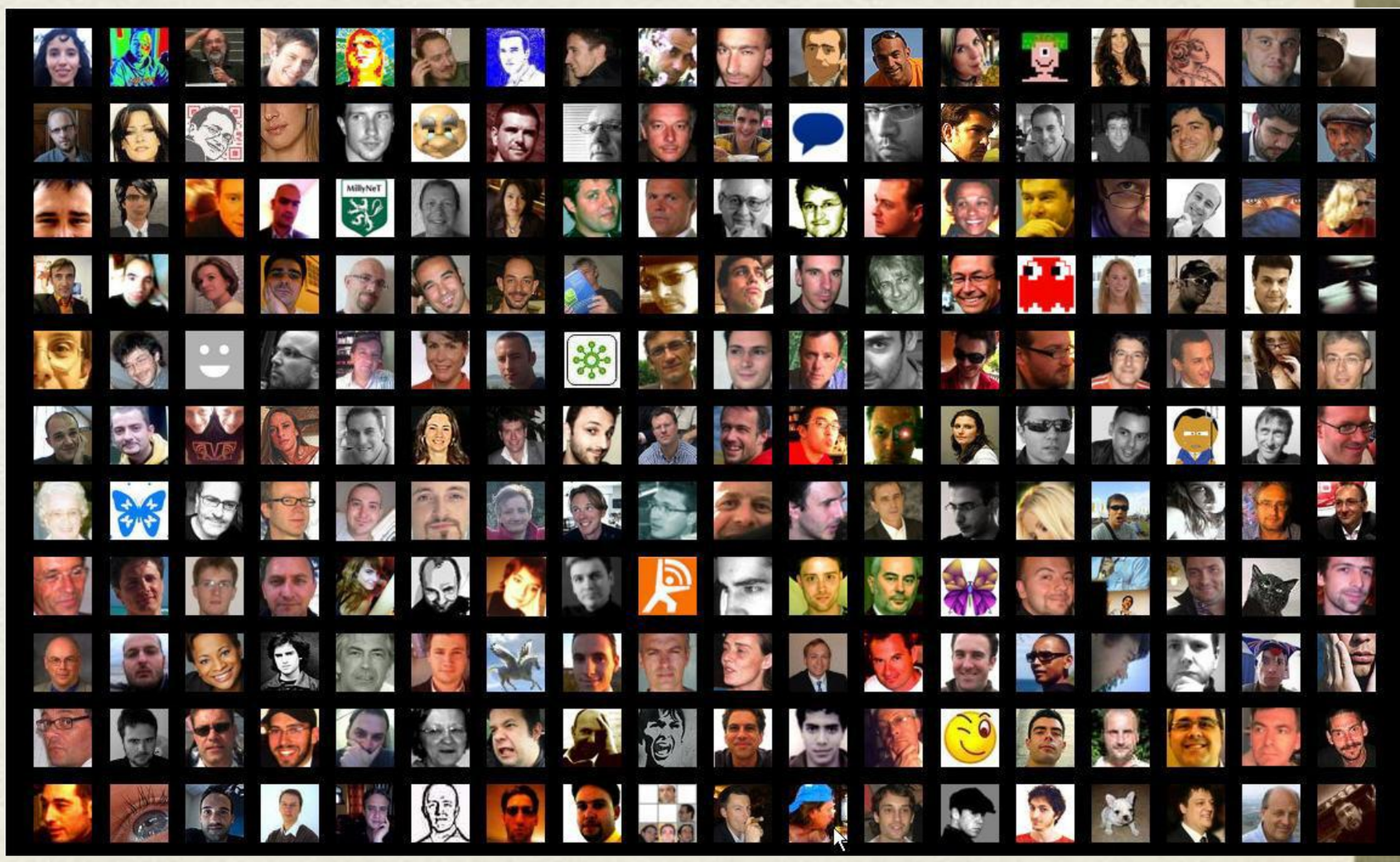
数据汇聚

数据海

实验数据

建设规划：整合海量资源

- 开放特藏库建设
- 开放微内容建设
- 开放资源建设
  - 文献数字化
  - 资源OCR
  - ……
- 从B2B走向C2B



浙江大学图书馆碑拓数据库

10.15.61.99:8080/btmain/

碑拓數據庫

2013年09月24日 星期二 10:55:46 | 首页 | 匿名用户 | [登录]

- 中國朝代表
- 金石類型
  - 摩崖
    - 摩崖
    - 刻石
    - 碣
  - 碑
    - 墓碑
    - 功德碑
    - 紀事碑
    - 題名碑
    - 宗教碑
    - 圖像碑
    - 書畫碑
  - 墓誌
  - 買地券， 鎮墓券
  - 造像碑 與 題記
  - 畫像石
  - 石經
  - 塔銘 與 經幢
  - 建築附刻 及 雜刻

選擇子庫類型: ☑ 歷代墓誌 ☑ 館藏拓片

選擇檢索方式: ◉ 簡單檢索 ○ 高級檢索

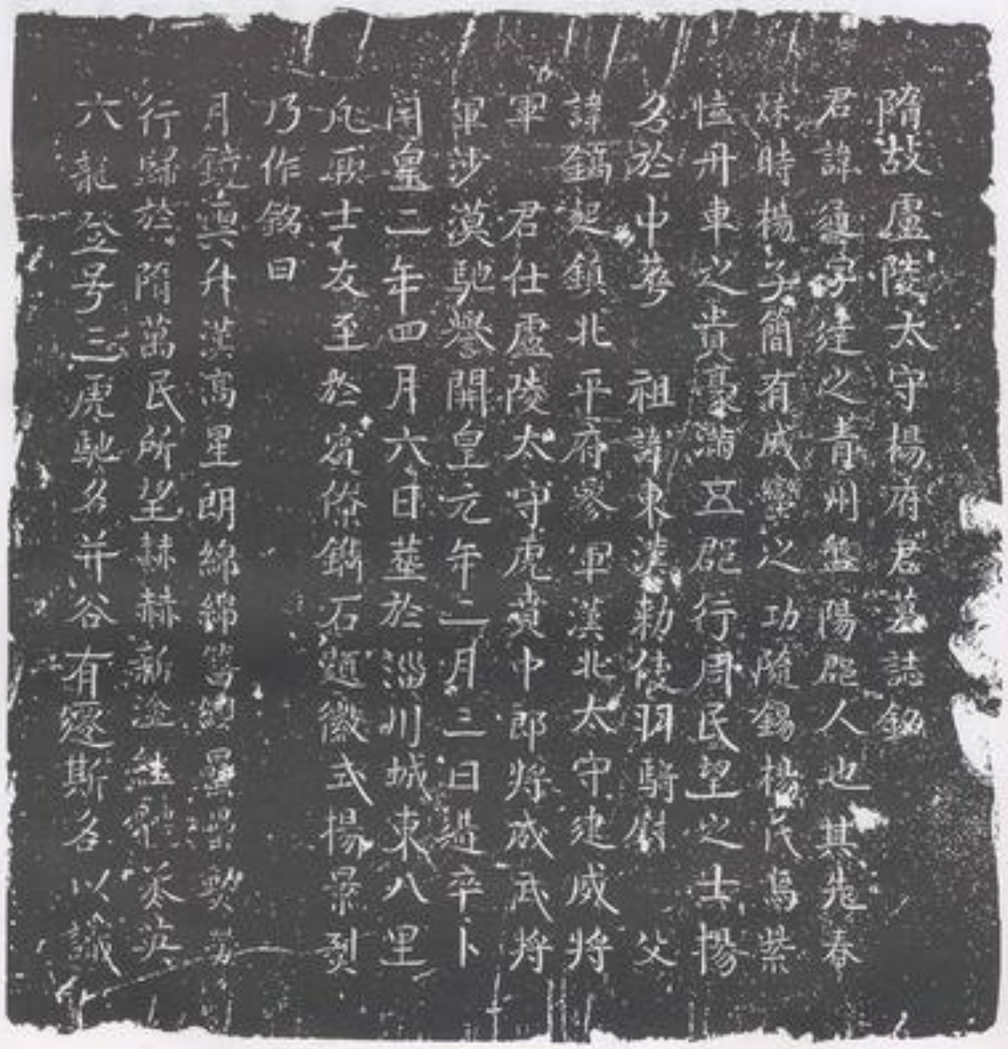
題名 | 任意匹配 | 請用繁體字檢索!

每次查詢最多返回: 全部 | 檢索 | 重置 | ☐ 二次檢索

序號	通用題名	收藏地點	朝代	紀年	金石類型	館藏地
1	梁暄誌	石存西安交通大學藝...	隋	開皇二年（公元五八...	墓誌	
2	高潭誌	石存河北省文物研究...	隋	北周大象二年（公元...	墓誌	
3	楊通誌		隋	開皇元年（公元五八...	墓誌	
4	北周武帝皇后阿史那...	石存成陽市文物保護...	隋	開皇二年（公元五八...	墓誌	
5	茹洪誌	石存西安碑林博物館...	隋	北周大象二年（公元...	墓誌	
6	楊元伯妻邵▬▬誌	曾歸長白端方、張仁...	隋	開皇二年（公元五八...	墓誌	
7	李和誌	石存西安碑林博物館...	隋	開皇二年（公元五八...	墓誌	
8	李君妻崔芷蘩誌	石歸河北正定劉秀峰...	隋	開皇二年（公元五八...	墓誌	
9	賀蘭祥妻劉氏誌	石存成陽市博物館。	隋	開皇二年（公元五八...	墓誌	
10	封子繪妻王楚英誌	石存中國國家博物館...	隋	開皇元年（公元五八...	墓誌	
11	張叔誌		隋	開皇三年（公元五八...	墓誌	
12	孫高誌	石存河南浚縣博物館...	隋	開皇三年（公元五八...	墓誌	
13	張顏誌	石存洛陽市考古所。	隋	開皇三年（公元五八...	墓誌	
14	寇熾妻姜敬親誌	石存西安碑林博物館...	隋	開皇元年（公元五八...	墓誌	
		曾存河南圖書館，今...	隋	開皇二年（公元五八...	墓誌	

建设规划：整合海量资源

- 开放微内容建设
  - 注册用户可以可以对图书进行：
    - 标签、标引
    - 书评、推荐
    - 笔记、注释等微创作
- 用户对微创作内容可以选择私有和开放
- 用户选择“开放”的内容经过CADAL平台审核后成为书籍内容的组成



1

我有图片☝

编辑 纠错

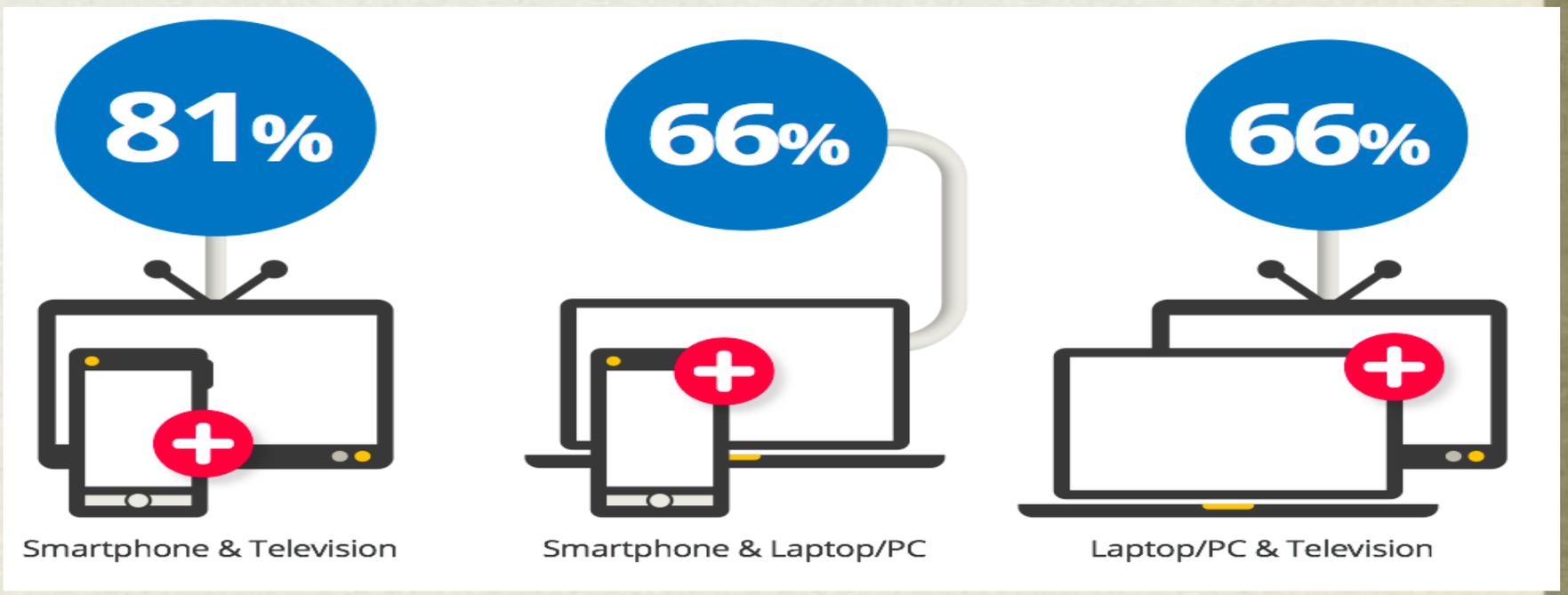
项目名称	子项目	项目内容
題名	通用名	楊通誌
	額題	隋故盧陵太守楊府君墓誌銘。
金石所在地	出土地點	山東淄博出土。
金石年代	朝代	隋
	紀年	開皇元年（公元五八一）二月三日卒，二年（公元五八二）四月六日葬。
金石類型	金石類型	墓誌
書體行款	銘文行款	誌文一四行，滿行一六字，正書。
拓片形態	尺寸	長四三、寬四一釐米。
附注	金石附注	其裨案：楊通及其祖、父均不見於正史。誌文『盧陵』原誤作『盧陵』。『紫▬』蓋卽『紫蓋』。『鎮北平府』，可補北朝鎮府名目之闕。國家圖書館藏拓本钤『金石臣之家學』、『耕石齋藏』印。
	錄文	隋故盧陵太守楊府君墓誌銘君諱通，字達之，青州盤陽郡人也。其先春秋時，楊子簡有滅蠻之功，隨錫楊氏焉。紫▬丹車之貴，豪滿五郡；行周民望之士，揚名於中華。祖諱東漢，敕使、羽騎尉；父諱鎬，起鎮北平府參軍、漢北太守、建威將軍。君仕盧陵太守、虎賁中郎將、威武將軍。沙漠馳譽，開皇元年二月三日遘卒。卜開皇二年四月六日葬於淄川城東八里。凡厥士友，至於賓僚，鐫石題徽，式揚景烈，乃作銘曰：月鏡雲升，漢高星朗。綿綿簪

建设规划：整合海量资源

- 开放资源建设
  - 文献数字化
    - 用户对于知识库缺藏内容提供数字化资源
    - 用户依据自身藏书提供数字化资源
    - 用户提供原创资源（课堂笔记、讲座视频……）
  - 资源OCR
    - 用户对CADAL文献提供纯文本资源
    - 用户完成10页文本校对即可以获得该书纯文本资源
    - CADAL资源用于全文搜索及资源发现服务

建设规划：泛在个性服务

- 数字图书馆与个人书房融合
  - CADAL云服务与读者私有云：注册用户可以拥有存储空间以构建自己的在线书房
- PC端、移动端一体化：多屏同一



建设规划：泛在个性服务

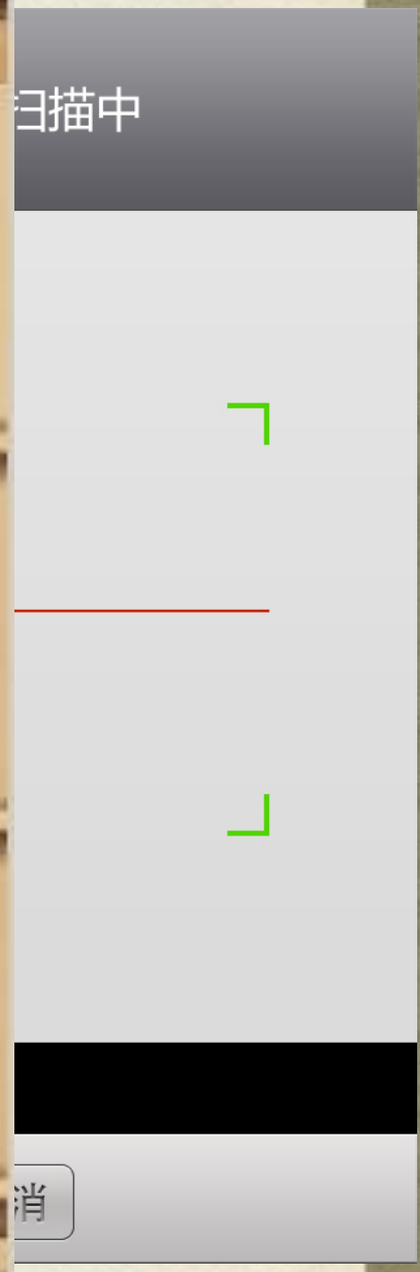
- 数字图书馆服务一体化：以用户为中心的个人书房
  - 个人资源获取
    - 搜索、借阅、收藏、购买、推荐、交换
  - 个人资源管理
    - 个人书房、CADAL书架、图书馆（可借）、书店（可买）
  - 个人内容创建
    - 书评、书单、标签、博客……
  - 个人研究平台
    - 时空坐标：关联展示
    - 文献索引：交互标注
    - 资源汇聚：众包协作
    - 知识管理：内容整合

建设规划：泛在个性服务


Bell
4:20 PM
CADAL


返回


Bell
4:20 PM
未登录
主页
拥有的书 27
想借的书 122
想买的书 122
出借的书 15
书店查询
设置
搜索
水浒传


扫描中
消

书名: 中国北朝石
精品集（套
作者: 李仁清

出版社: 大象出版社

简介

目录

收藏状态

本书收藏在 <想买的书>

加入我的书架

书名: 西北民
馆于右
作者: 郭郁烈
版社

出版社: 上海古

简介

目录

标签

金石拓片

收藏状态

本书收藏在 <想买的书>

加入我的书架

金石拓片

添加标签

收藏状态

本书收藏在 <想买的书>

加入我的书架

我要借

馆藏状态:

卷期	单册状态 | 应还日期	分馆 | 馆藏地	架位	扩展	预约	预约数	条码
	图书阅览 | 在架上	院系分馆-古代文学 |	K877.22/CG1	扩展			00000475230

我要买

亚马逊 (RMB228.60)
京东商城 (RMB228.60)
99网上书城 (RMB238.40)

返回 书目详情 编辑

书名: 礼品装家庭必读书: 全真图解本...

作者: 李时珍

出版社: 辽海出版社

简介

目录

收藏状态

本书收藏在 <拥有的书->中医>

出借

请输入借阅人

出借

借阅电子书 编辑标签

拥有的书

人文社科 (6)

中医 (1)

艺术 (1)

扫一扫

趋势和预测？


netlash
bSeen
The best way
to predict
the future is
to invent it.
Alan Curtis Kay
Trendrapport 2012, de experts aan het woord.
Trendrapport 2012, de experts aan het woord.

生存法则


There will be only 2 types of companies left.
The **quick** and the **dead.**


Thank You


Questions?