華東師範大學
EAST CHINA NORMAL UNIVERSITY

图书馆智慧数据实践

华东师范大学 许鑫

2021-07-02

華東師範大學
EAST CHINA NORMAL UNIVERSITY

图书馆智慧数据

1 资源服务之变

◆数据浪潮颠覆传统

◆资源走向智慧数据

2 数据智慧之路

◆数据智慧化路漫漫

◆数据智慧化需躬行

1 资源服务之变

全球的数据管理政策

《柏林宣言》（2003年）

明确将科研数据作为学术知识的一部分

《全球科学信息共有倡议》（2005年）

倡议人们促进科学数据的合理传播和利用，并推出了服务于全球的数据管理系统UNDESA。

“联合共建数据公平港口”会议（2014）

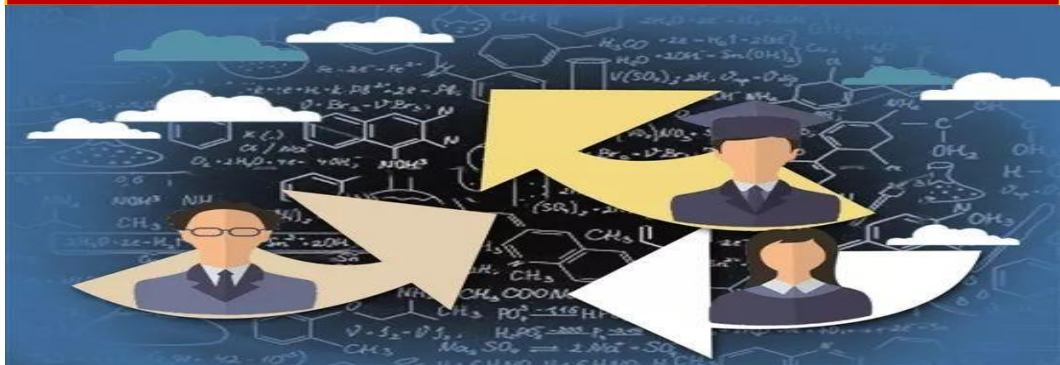
提出数据共享FAIR原则

“地平线2020”计划（2017）

要求所发表的研究论文必须开放出版或在出版后存储到开放知识库。

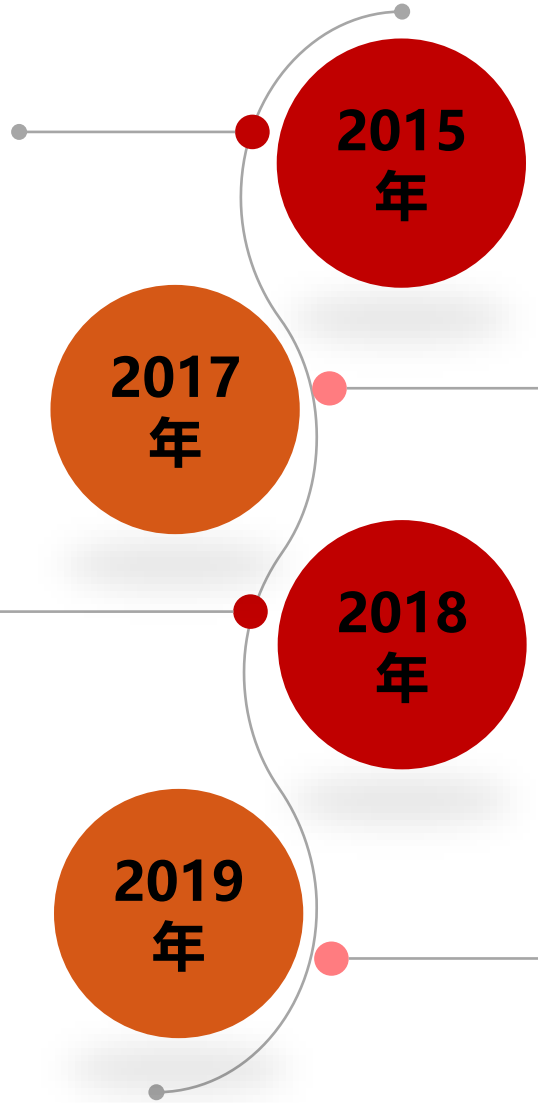
世界各国在数据管理方面的推进措施

全球各个国家和地区政府意识到了数据管理与共享的重要性，纷纷推出自己的数据管理政策与相关系统。



《促进大数据发展行动纲要》

加快政府数据开放共享，**推动资源整合，提升治理能力**；推动产业创新发展，培育新兴业态，助力经济转型；强化安全保障，提高管理水平，促进健康发展。



《大数据产业发展规划（2016-2020年）》

围绕实施国家大数据战略，以强化大数据产业创新发展能力为核心，以**推动数据开放与共享**、加强技术产品研发、深化应用创新为重点，以完善发展环境和提升安全保障能力为支撑，打造数据、技术、应用与安全协同发展的自主产业生态体系，全面提升我国大数据的资源掌控能力、技术支撑能力和价值挖掘能力，加快建设数据强国，有力支撑制造强国和网络强国建设。

《科学数据管理办法》

加强和规范科学数据管理，要适应大数据发展形势，积极推进**科学数据资源开发利用和开放共享**，加强重要数据基础设施安全保护，依法确定数据安全等级和开放条件，**建立数据共享和对外交流的安全审查机制**，为政府决策、公共安全、国防建设、科学研究提供有力支撑。

《关于进一步弘扬科学家精神加强作风和学风建设的意见》

论文等科研成果发表后1个月内，要将**所涉及的实验记录、实验数据等原始数据资料交所在单位统一管理**、留存备查。

要运用大数据提升国家治理现代化水平，善于获取数据、分析数据、运用数据，是领导干部做好工作的基本功！

——习近平,2018

◆ 数据浪潮颠覆传统：“变”

研究范式转变之重数据

第一范式—经验范式

由范式。伽利略、哥白尼以及同时代的开普勒创建的实验观察模式，被称为科研第一。

第二范式-模型推演和理论科学

以牛顿微积分和经典力学为代表的模型推演和理论精准预测，是科研的第二范式。

第三范式-仿真模拟和计算科学

20世纪初，量子力学和混沌理论的发展否定了模型推理和理论预测的可行性，并以电子计算机的诞生为契机，演变出科研的第三范式——计算科学。

第四范式-数据密集型科学

随着小世界网络和无尺度网络等复杂网络研究的深入，以及计算能力和传感器的无处不在，数据密集型科学从计算科学中分离出来，成为科学研究的第四范式。

几千年前　几百年前　几十年前　当今

“数据密集型”科学研究——指当今科学研究越来越依赖于数据的聚集和分析，特别是海量数据分析。

数据驱动研究范式的兴起

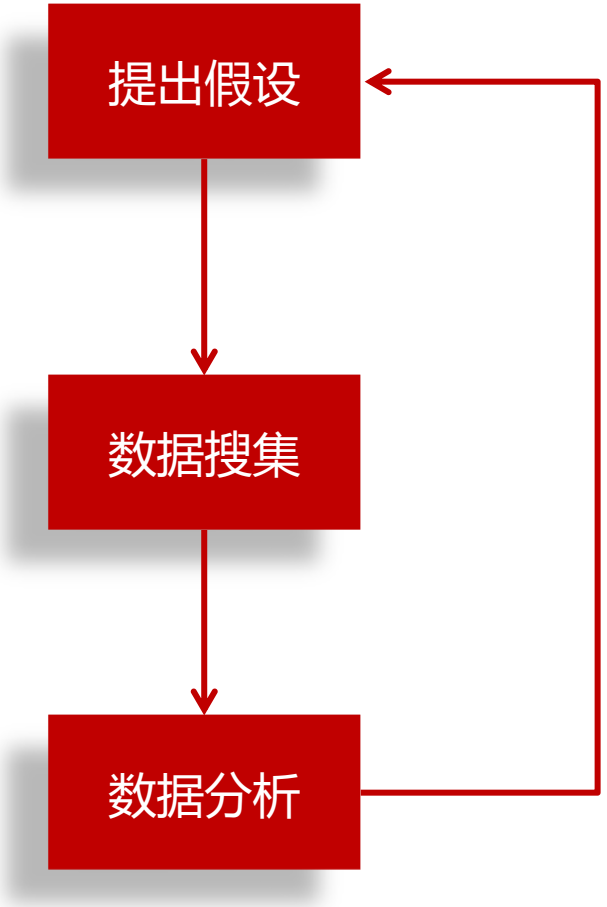
自上而下实证研究范式转向自下而上的数据驱动范式

传统的实证研究：自上而下的决策和思维过程

强调在理论的前提下建立假设，收集数据，证伪理论的适用性，采用随机抽样的定量调查问卷获取数据，验证假设，你不问的问题被访者也不会回答。

数据驱动：自下而上的知识发现过程

重在发现知识，预知未来，为探索未知的社会现象带来机遇。这种预见性是一种自下而上的知识发现过程，是在没有理论假设的前提下去预知社会和洞察商业趋势、规律。



研究视角转移之重内容、关系

通过相关关系算法**分析**海量的数据，从而做出**预测**是大数据时代的核心。

内容：大数据的大数量提供了新的研究空间，所涉及的信息、数据远远超过一般的阅读、分析和理解所能处理的范畴，是以往“不可研究”或“难以研究”的内容。

关系：对其相关关系做出的结论并非观察、思索、领悟等传统方法获得，而是通过对大量数据引入计算分析方法，而“自动涌现”。

数据挖掘（data mining）、文本挖掘（text mining）、网络挖掘（web mining）、自然语言处理、机器学习……

技术手段发展之重融合、计算

- **多源异质网络融合技术**

  丰富的数据获取环境，多样的数据种类结构，跨平台异质网络，相对以往单一的数据采集处理技术发生了较大的改变。

- **高性能并行计算技术**

  云平台的搭建、深度学习算法设计，在海量数据环境下，以往针对单机模式的算法和处理流程不在适用，高性能并发实时处理技术应运而生。

社会需求之新

大数据逐渐在行业开始深度应用

人口大数据、社会治理大数据、大数据招商、大数据扶贫

社会对研究的需求逐渐由理论研究转向数据研究

需要研究者借助数据分析的手段发现更多的问题，通过数据的梳理和挖掘寻找解决的方法。

数据的快速发展对数据整合和共享提出了现实要求

比如：上海已建成人口库、机构法人库、还在进一步建设征信数据库，信息组织的主体相对完善，但是缺乏整合和共享机制，如果实现数据价值的爆发性增长，为应用学科的研究者提出了数据驱动的现实需求。

研究领域之新

- **社会计算**

  广义而言，社会计算是指面向社会科学的计算理论和方法；狭义而言，社会计算是面向社会活动、社会过程、社会结构、社会组织及其作用和效应的计算理论和方法。

- **数字人文**

  数字人文受到了历史学、文学、语言学、图书情报学、艺术学等多个学科学者的关注，催生了众多领域导向性专题数据库建设与探索性研究项目，是人文研究在未来数字空间中延续繁荣的基础。

◆资源走向智慧数据

数字化

将图书馆中不同的信息资源以**编码**的形式进行加工和存储，以方便用户进行检索和利用。其只是实现馆藏资源**向数字世界的映射**，能够被计算机所存储、处理和展示，仅仅转换了一种存在方式而已。

信息时代

数 据 时 代

是按照**知识单元**的方式来组织领域知识，从而能够构造一个模拟领域应用的知识环境。其是实现**对资源语义内容的获取与管理利用**，是数字化的下一阶段。

数据化

智慧数据

通过对具有语义关联的元数据进行融合、分析等活动，使其自带智慧，能够**刺激创新和服务**的“聪明数据(Smart Data)”。

2 数据智慧之路

◆数据智慧化路漫漫！

数据赋智之难

现状：

- 数字化全文扫描
- 简单元数据加工
- 各独立的资源数据库建设

问题：

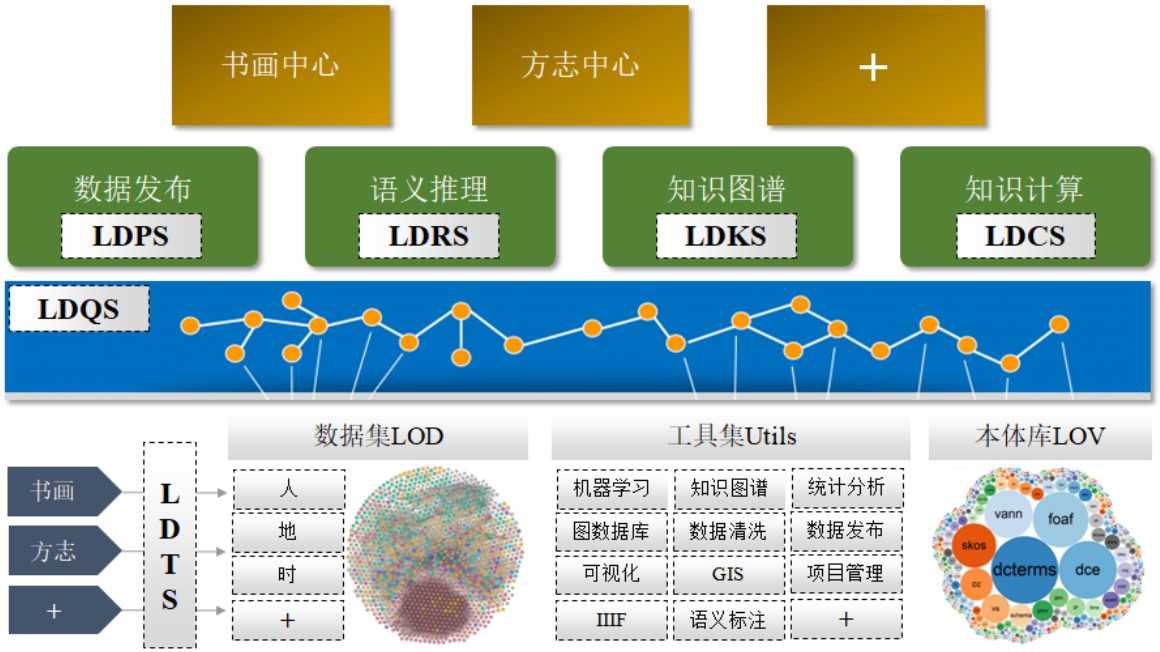
- 图书馆内部：专题专项的建设模式，资源多源、异构，开放性差且重复建设。
- 图书馆之间：各个数据库之间粘合度和共享度不高，各馆的同类知识没有实现聚类和统一揭示，不能满足用户对某一学科知识全面性掌握的需求，不利于知识发现。

需要：

- 构建基于数据规范的基础设施，实现分散异构的**数字资源向机器可读、可理解、关联共享的优质数据转换**。
- 构建广泛参与的分布式平台，让各馆的数据在开放环境下按照一定的格式实现共享，用户**从大量的优质数据中发现新知的应用智慧**。

◆数 据 智 慧 化 需 躬 行

（1）基础设施：语义支撑平台



应用平台层

作为资源与用户交互的直接入口，为用户提供更为精准的分析数据，助力科学研究。

知识服务层

可以结合本体、知识图谱和机器学习方面的理论和算法，提供资源语义相关的知识服务。

语义关联层

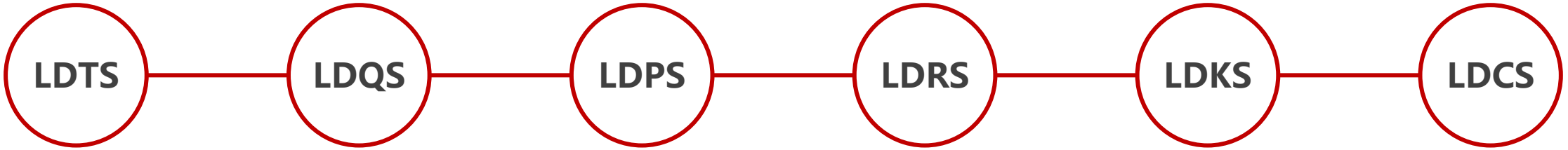
对同类资源的不同本体进行对齐，建立不同资源之间的语义关联，实现资源语义层面的统一。

基础设施层

开源工具和开放本体，将多源异构的数据转换为RDF结构的数据，实现资源语法层面的统一。

语义支撑平台：数字资源向关联数据迈进

六大关联服务



关联数据转换服务

为语义关联层提供标准数据支持

关联数据检索服务

为知识服务层提供链接的语义数据

关联数据发布服务

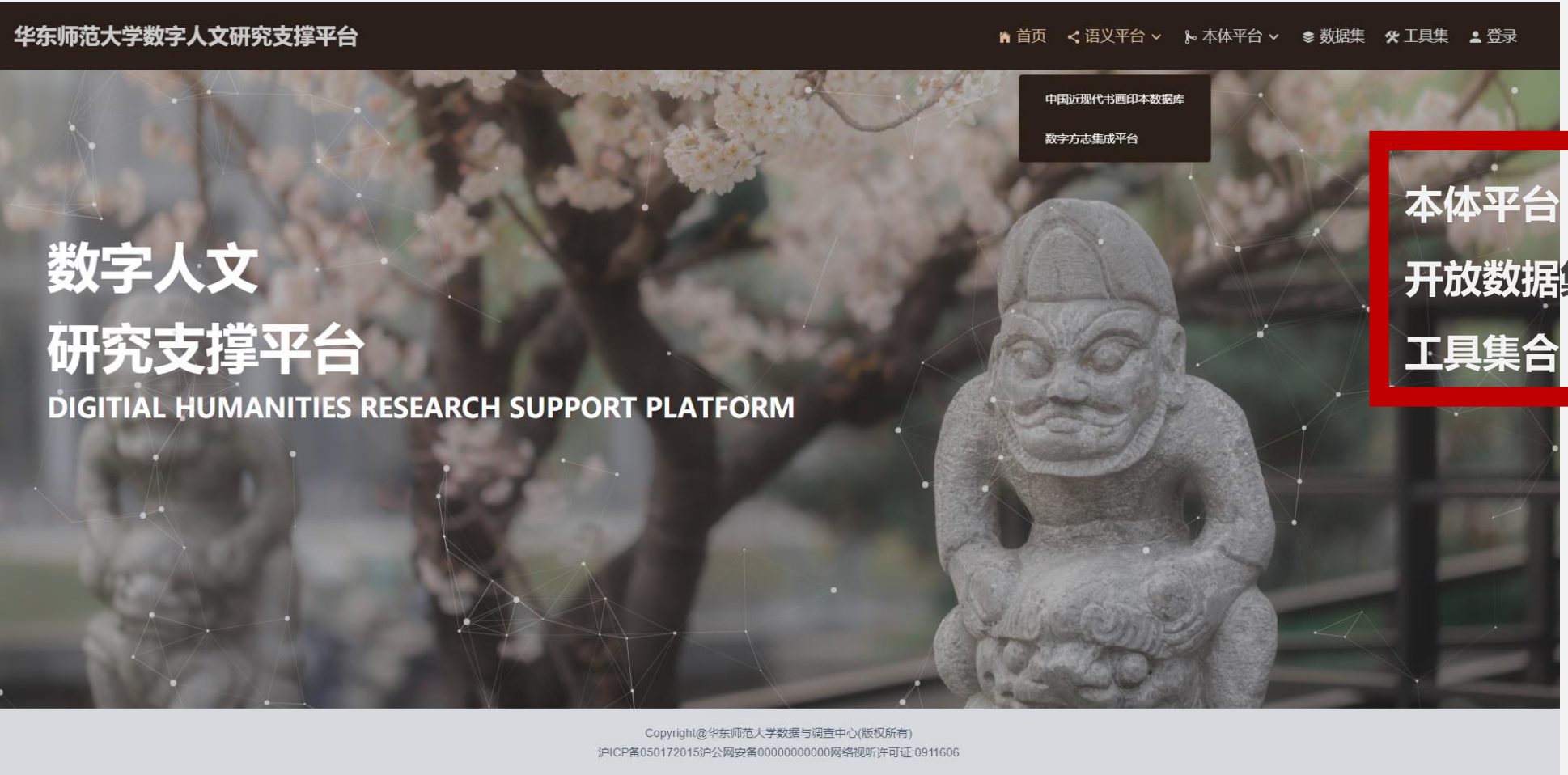
关联数据推理服务

关联数据知识服务

关联数据计算服务

是整个语义平台的价值体现，为最终的应用平台层提供方法和手段

华东师范大学数字人文研究支撑平台



本体平台

开放数据集

工具集合

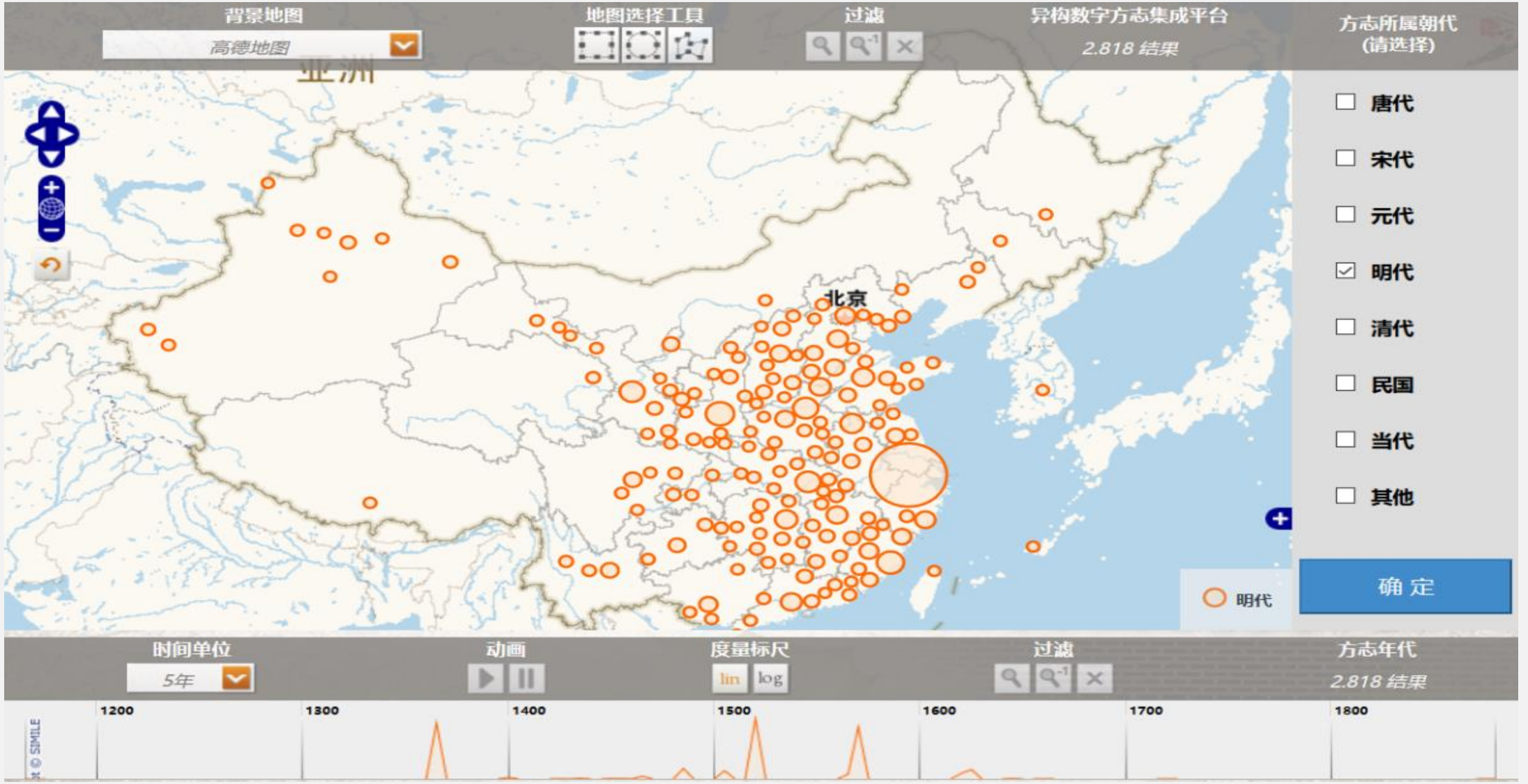
书画中心

数字方志集成平台

华东师范大学数字人文研究支撑平台

实践项目	LDTS	LDQS	LDPS	LDRS	LDKS	LDCS
书画中心	√	√	√		√	√
方志平台	√	√	√		√	

方志平台GIS时空分析



使用MARKUS对方志内容进行自动标记

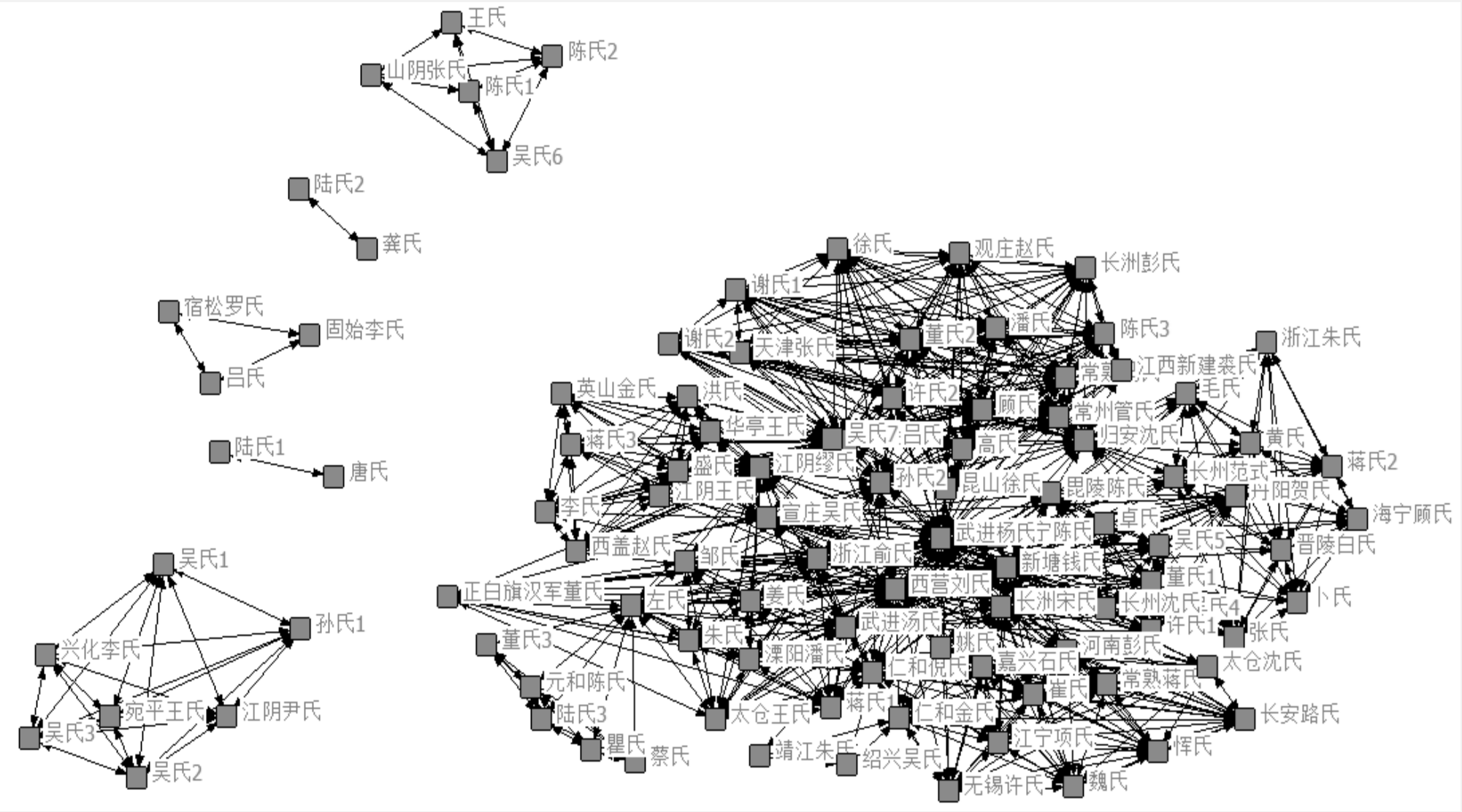
书画中心IIIF


IIIF
添加书签
更改布局
全屏
董香光山水冊集外名品第一
Index
Search
Layers
封面
Page 001
Page 002
Page 003
Page 004
Page 005
Page 006
Page 007
Page 008
Page 009
Page 010
Page 011
董香光山水冊
中國名畫集外冊第一
有正書局印行
精一文献

民国报纸广告图像资料语义标注

以民国报刊《新闻报》上的越剧广告为研究对象，结合语义模型以及IIIF技术支持的平台，准确地揭示图像所涵盖的文本信息

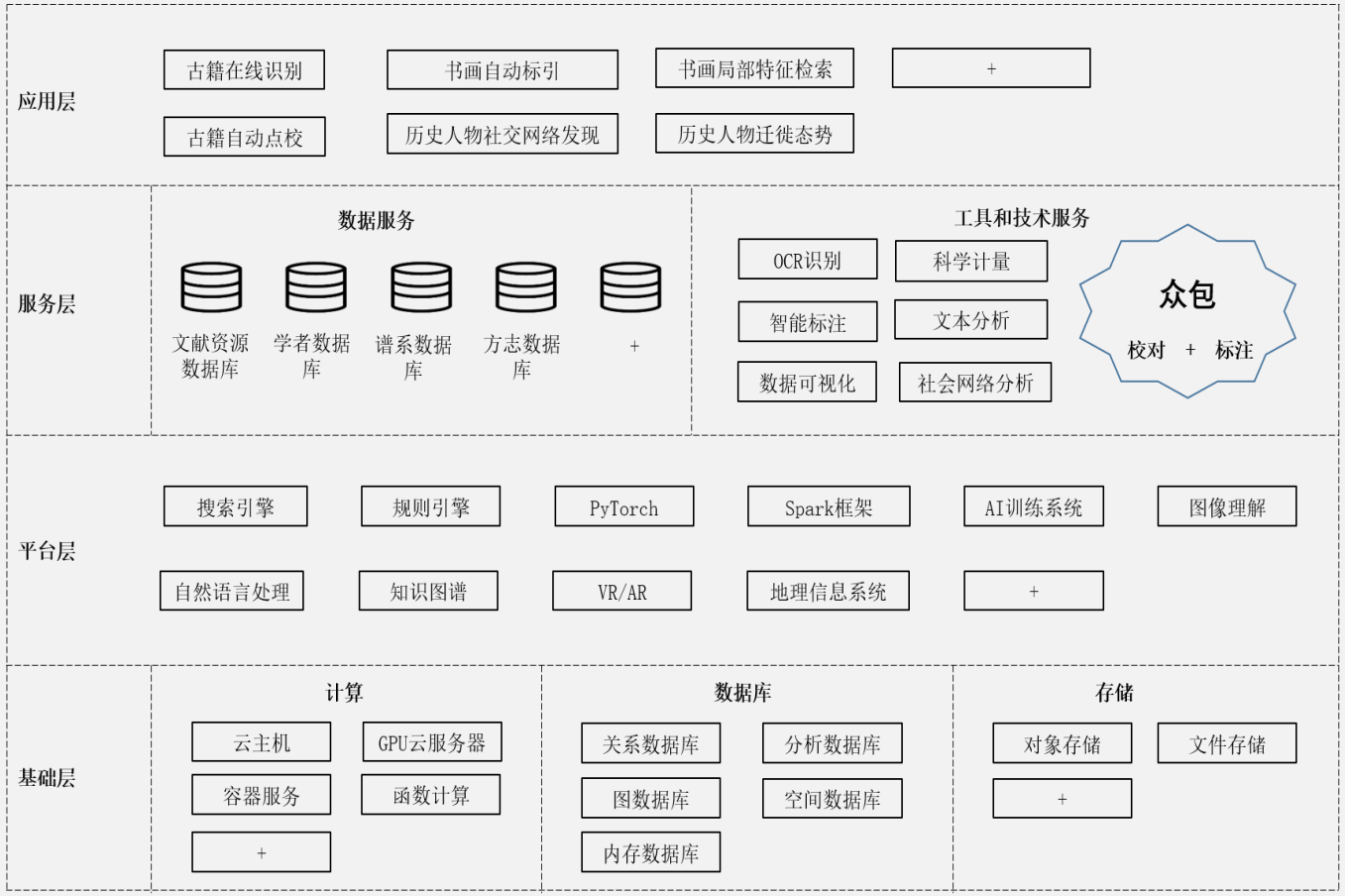
世家姻娅社会网络可视化



世家姻娅交往网络

◆数 据 智 慧 化 需 躬 行

（2）基础设施：人工智能平台



应用层

将复杂的工具和技术以友好、便于利用的界面呈现给人文研究者

服务层

数据服务、工具和技术服务

平台层

通用的人工智能框架和中间件服务

基础层

计算、数据库和存储

华东师范大学数字人文人工智能平台

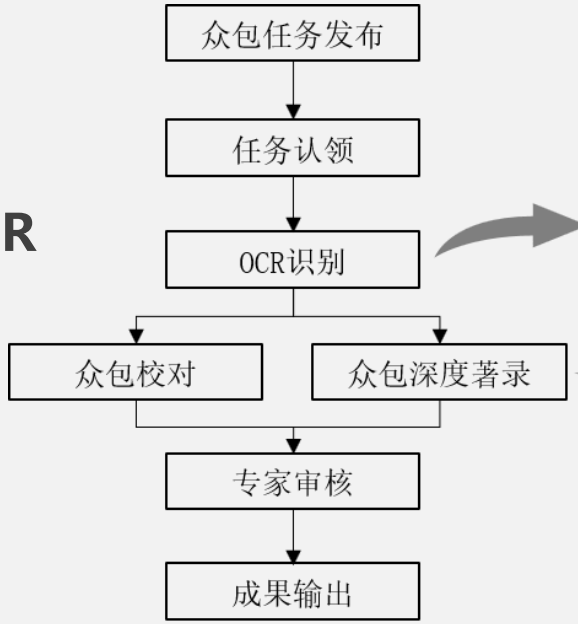
V3.0：图文识别 + 众包

V2.0：抄本识别 + 句读

V1.0：刻本识别 + 夹注

华东师范大学数字人文人工智能平台

融合机器学习的OCR



众包管理机制

校对　　深度著录

◆数据智慧化路漫漫II

数据管理之难

现状：

- 理念：围绕数据本身进行管理
- 实践：半开放单体系统模式

问题：

- 管理理念落后，基于数据本身进行管理，忽视用户需求。
- 传统的数据管理系统平台存在若干问题，难以支撑数据与知识服务。

需要：

- 构建**面向用户与服务**的新一代图书馆数据管理平台。

若干亟待解决的问题

数据库建设过程不规范，系统平台可用性不强

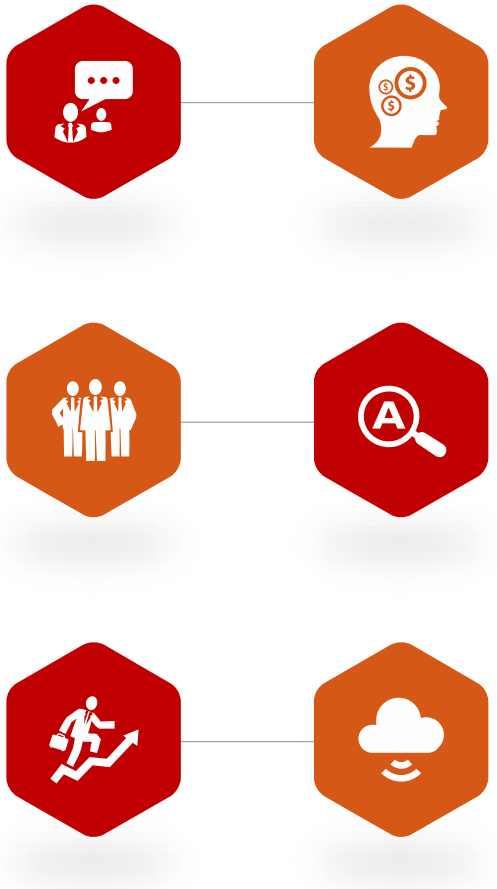
数据库系统选型落后，数据服务平台功能单一，检索效率低下，不支持机器读取和原始下载，系统平台整体上可用性较差，难以满足项目之外用户的实用性需求。

数据质量难以控制，内容可信度无法保障

在数据采集发布过程中，缺少质量控制标准和规范，又无问责机制，导致数据的可信性与可靠性得不到保障。

数据共享分散，获取难度大

良性的开放共享文化和机制尚未形成，数据分散在不同系统平台，缺乏统一的管理与共享，用户之间寻找和获取数据的难度较大。



缺乏标准，数据移植性差

自建数据库为主，发布自由，标准不统一，数据机读能力差，可移植性差。

数据主权无法保障，贡献意愿低

数据被利用后，无法对利用的结果进行追踪，数据主权无法保障，降低了用户贡献数据的意愿和积极性。

缺乏分析工具，分析结果受多种因素影响

仅有传统少数的分析工具可用，面对全样本数据，缺乏大数据分析的平台和工具，并且受到用户自身计算机操作能力、分析工具等影响，分析结果质量层次不齐。

◆数 据 智 慧 化 需 躬 行

(3) 基础设施：数据管理平台

人文社科研究数据的特点

智慧数据

数据规模小，但蕴含的语义内容丰富多样。

数据分散，低水平重复建设较多

除了国家资助的重点项目，大多数数据都保留在科研工作者自己的手里，低水平重复建设较多，难以形成聚合效应。

高度可复用

人文社科数据的使用周期较长，同一研究方向的社科数据可以被多个研发团队复用，数据可以产生持续的价值。

面向全生命周期的数据管理平台

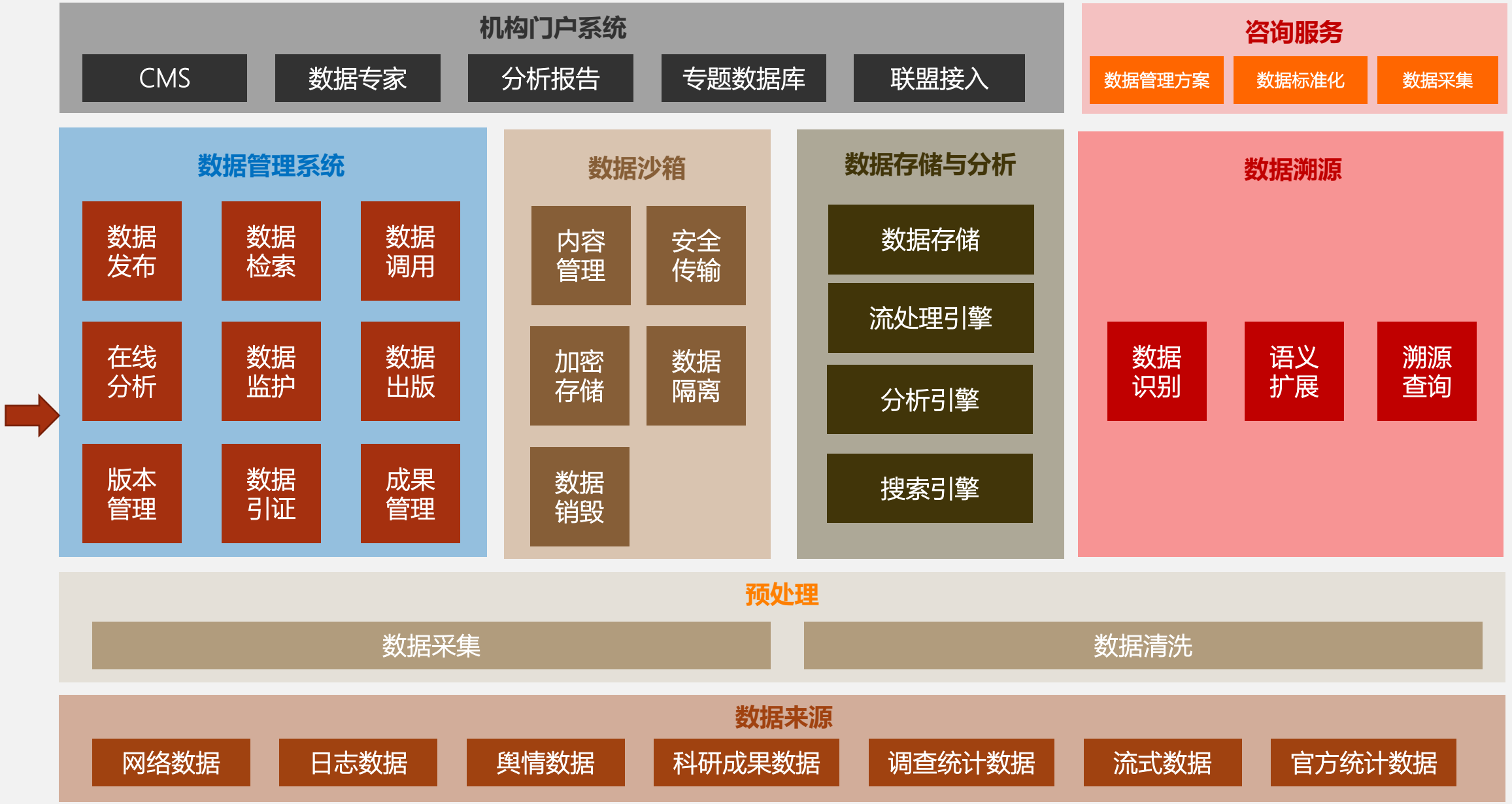
逻辑示意图

识别数据的演变阶段和用户的特定需求

实现可持续发展

面向全生命周期的研究数据管理平台

系统架构图



系统实例


华东师范大学 经济与管理学部
数据管理平台
主页
数据浏览
关于我们
帮助
登录
注册
开放共享 ○ 协作创新
以自建分享、自助提交等方式，广泛汇聚与共享各类科学数据，助力科学数据的重用和深度发现。
一站式检索
在数据条目中搜索
搜索
关键词搜索
新闻记者
二十四节气
民俗旅游
数据提交
数据服务
上传我的科研数据
提交我的数据需求

系统实例

0	6	3649
当日新增数据条目数	总数据空间数	总数据条目

精品数据

主题数据聚合

系统实例


华东师范大学 经济与管理学部 数据管理平台
主页
数据浏览
关于我们
帮助
zlyao@infor.ecnu.edu.cn
授权用户
数据发布者
2008-2019年图书情报与数字图书馆文献题录数据
杨佳颖
学科
图书馆、情报与档案管
发布时间
2019-10-30
原始出处
中国知网
类型
数据集
语言
中文
展开
上传文件
请填写申请原因
请输入申请原因
仅数据发布者，拥有数据开放的权限
不能为空
取消
确定
文件
操作
2008-2019年图书情报与数字图书馆题录数据.csv
暂无
42.72MB
申请访问

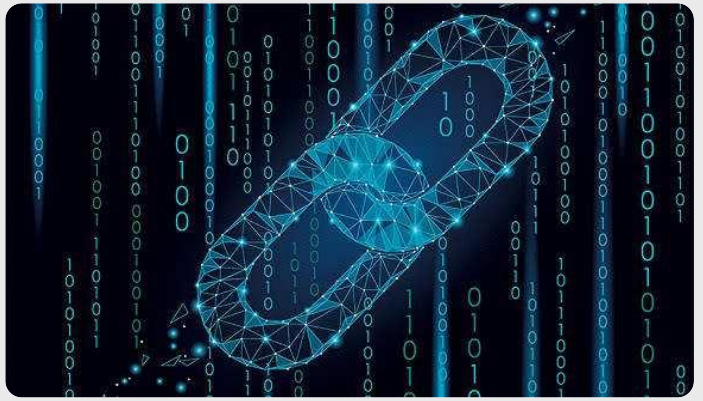
系统实例


华东师范大学 经济与管理学部 数据管理平台
主页
数据浏览
关于我们
帮助
zlyao@infor.ecnu.edu.cn
首页 / 个人中心 / 申请授权管理
数据空间管理
数据集合管理
数据条目管理
申请授权管理
基本信息
修改密码
申请/审批管理
我的申请
我的授权
未获批的数据申请
申请文件
所属条目
申请的间
申请原因
状态
2008-2019年图书情报与数字图书馆题录数据.csv
2008-2009年图书情报与数字图书馆文献题录数据
2019-10-30
研究需要，特此申请。
待审核
1

社会科学数据共享平台

V2.0

校际联盟



纵跨三个层次进行社科领域数据共享

V1.0

本地部署+校内共享

V3.0

全球公链

数据共享联盟

政府管理部门

- 数据资源格式规范
- 数据资源建设规范
- 数据资源运维规范
- 数据资源评估标准
- 数据资源汇交机制
- 数据资源管控机制
- 数据资源开放机制
- 数据资源专题服务

出版机构，图书馆

- 数据出版机制
- 数据生产机制

数据共享联盟

大数据基础设施厂商

数据分析与应用厂商

- 数据在线分析平台
- 微服务容器
- 数据溯源系统
- 数据评价系统
- 数据收割
- 数据存储
- 数据监护
- 数据发布
- 数据调用
- 数据分析
- 元数据平台
- 云存储平台
- 区块链设施

大数据处理与服务厂商

数据库相关厂商

联盟成员单位

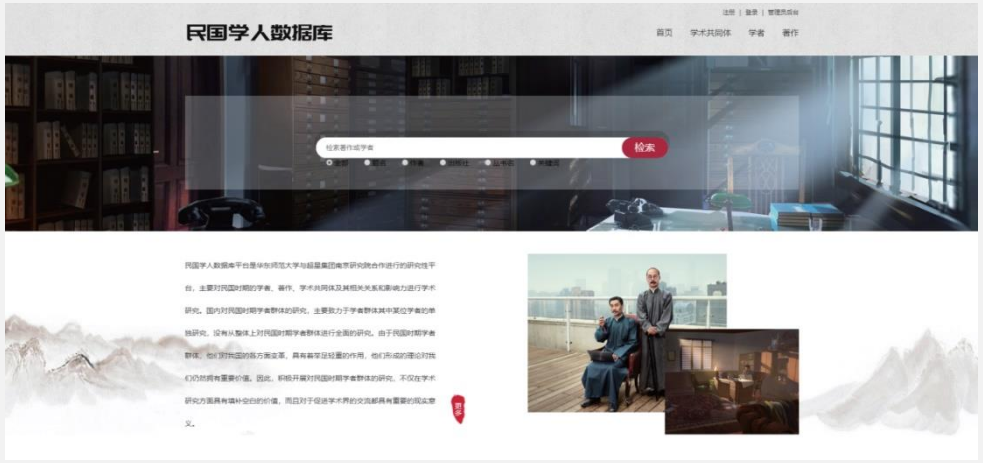
- 高等院校
- 社科院系统
- 公共文化机构
- 政府政策单位
- ……

◆数据智慧化需躬行

专题数据库

江南文化文献资料数据库

世界中国学研究者专题数据库

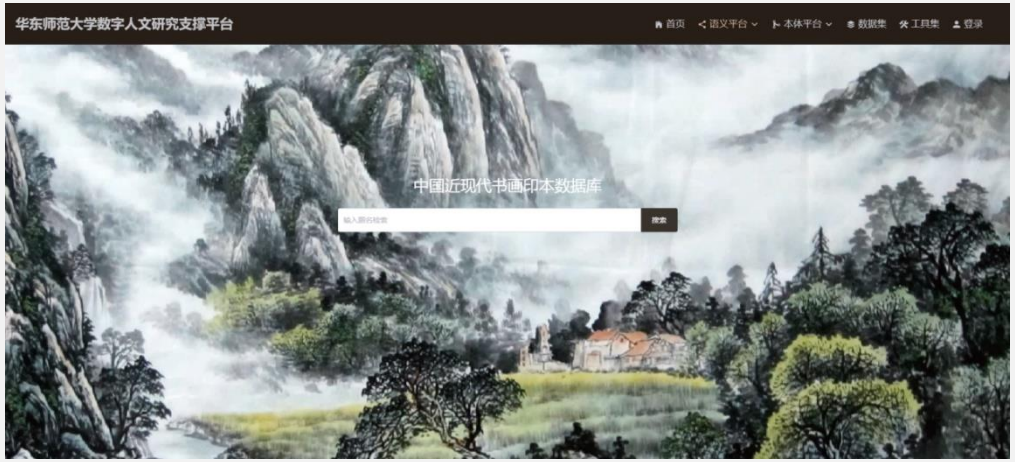


民国学人专题数据库

华人华侨学者专题数据库

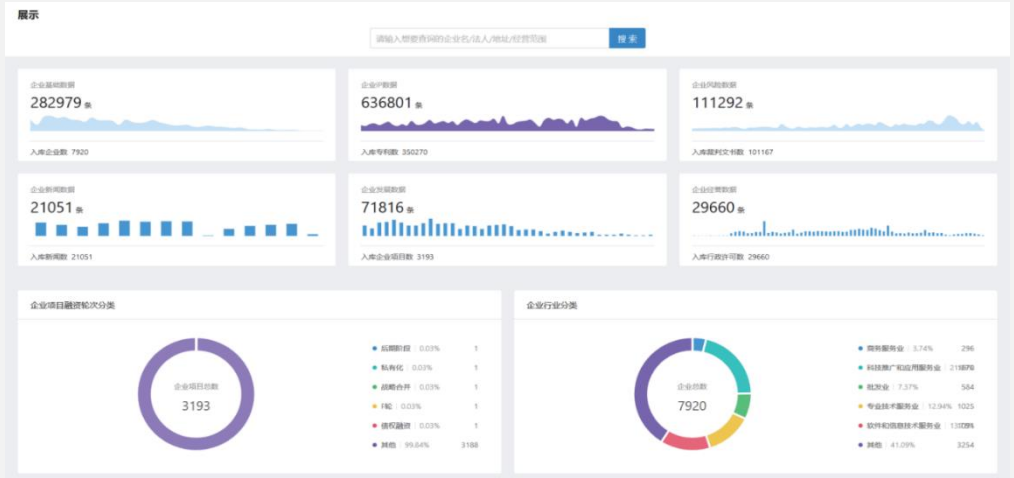
◆数据智慧化需躬行

专题数据库



中国近现代书画印本数据库

中小学教师专题数据库



上海高新技术企业数据库

创新创业开放数据平台

華東師範大學
EAST CHINA NORMAL UNIVERSITY

谢谢各位！