DQMiS

第二届数据质量管理国际峰会

The 2nd Data Quality Management International summit

北京大学 PEKING UNIVERSITY

高校学科数据平台的构建与应用

主讲人：王继民

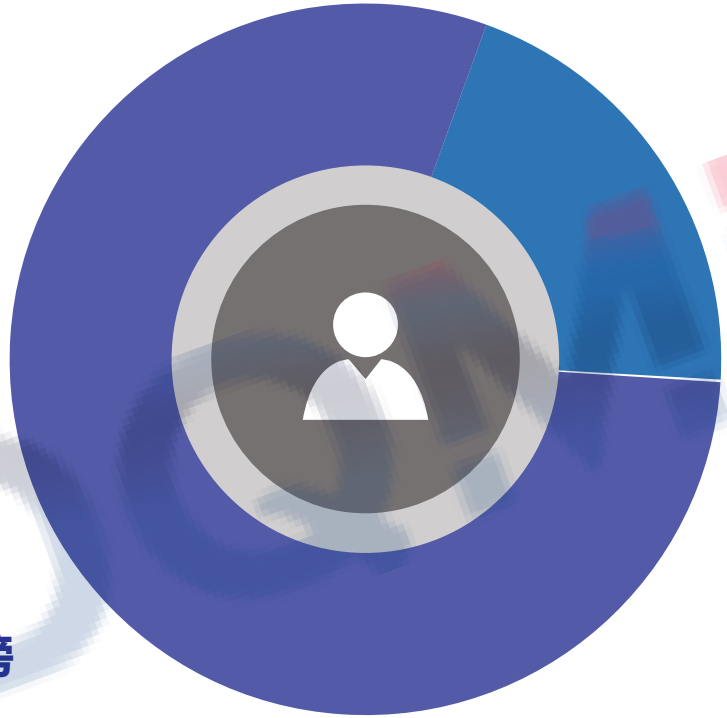
2018年9月

目录
CONTENTS

01

研究背景

国内外大学与学科评价---两类评价



内部评价（自评）：科研机构（或学科）的内部评价

- 中国科学院对所属“研究所”的评价、同济大学的院系评价等

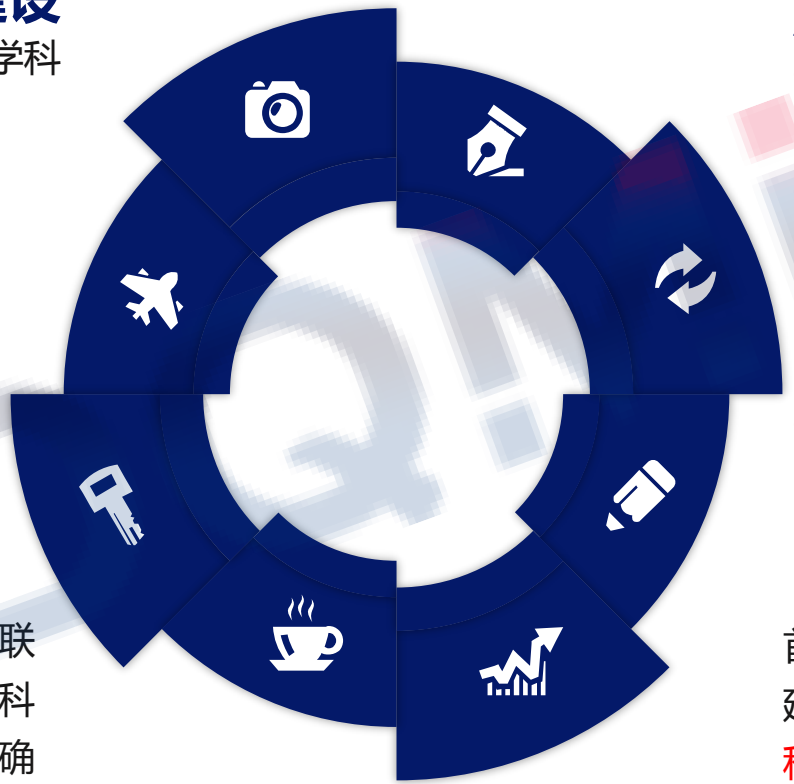
外部评价：各类大学或学科排行榜

- QS世界大学（学科）排行榜、《泰晤士高等教育》、《美国新闻与世界报道》、上海交大等
- 教育部的一级学科评估、武汉大学评价研究中心等

“双一流”建设

建设

世界一流大学和一流学科



总体目标

- 到2020年，若干所大学和一批学科进入世界一流行列，若干学科进入世界一流学科前列
- 到2030年，更多的大学和学科进入世界一流行列，若干所大学进入世界一流大学前列，一批学科进入世界一流学科前列，高等教育整体实力显著提升
- 到21世纪中叶，一流大学和一流学科的数量和实力进入世界前列，基本建成高等教育强国

2017年9月，教育部、财政部、国家改委联合发布《关于公布世界一流大学和一流学科建设高校及建设学科名单的通知》，正式确认公布世界一流大学和一流学科建设高校及学科名单

首批双一流建设高校共计137所，其中世界一流大学建设高校42所（A类36所，B类6所），世界一流学科建设高校95所；双一流建设学科共计465个

外部评价：各类大学或学科排行榜

目前国际上至少有40多个知名的大学（或学科）排行榜，并有增多的趋势。为吸引留学生，2010年欧盟开始研究大学排名

- QS世界大学（学科）排行榜
- 英国的《泰晤士报高等教育副刊》
- 美国《美国新闻与世界报道》
- 德国的《明镜》周刊、德国的高等教育中心
- 加拿大的《麦肯林》
- 日本的《朝日新闻》、印度《Data quest》
- 俄罗斯的《职业》、 西班牙：世界大学网络排名
- 香港教育评审局、台湾大学高等教育评鉴中心、淡江大学
- 上海交通大学、武汉大学、广州管理科学研究院、中国校友会、网大、浙江大学等15个单位进行研究发布

关于评估结果

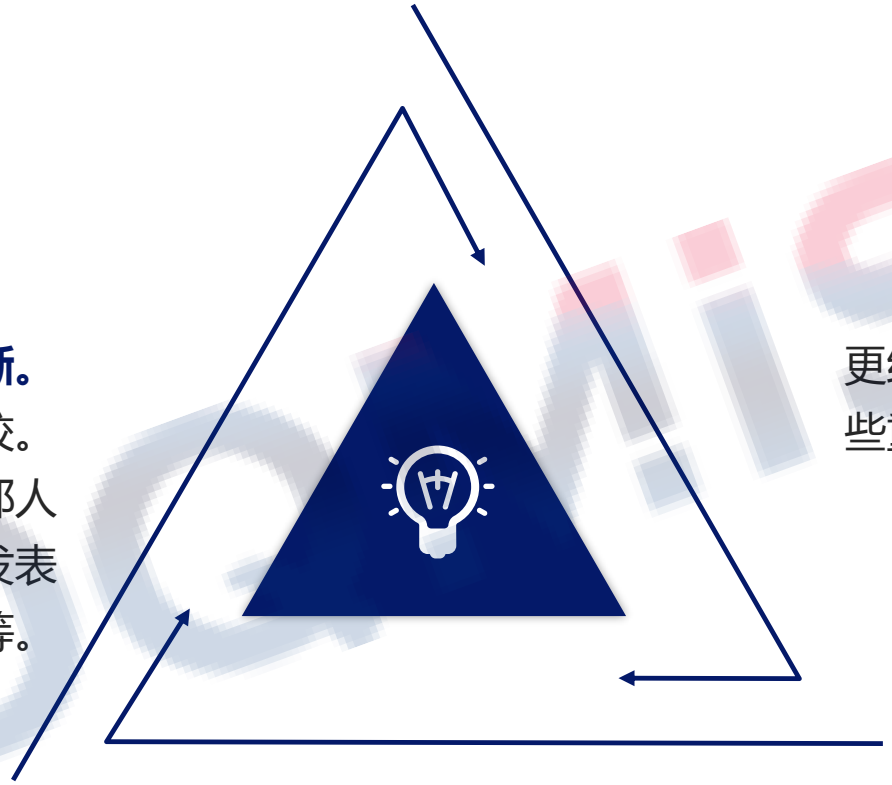
- 大多数评估只公布名次或总得分，最多也只会公布一级指标的得分情况，再详细一点的数据不得而知

教育部第三次学科评估 “图书情报与档案学”一级学科	
武汉大学	96
南京大学	86
中国人民大学	85
北京大学	79
华中师范大学	76
中山大学	
南开大学	74
吉林大学	

教育部第四次学科评估 “图书情报与档案学”一级学科	
A+	南京大学
	武汉大学
A-	中国人民大学
B+	北京大学
	南开大学
	华中师范大学
	中山大学

关于评估结果---以图书、情报与档案学为例

具体那些指标强、那些弱并不清晰。
如：武大、人大、南大等兄弟院校。近几年获得国家社科基金、教育部人文社科的情况（数量与数额），发表论文情况（国内外），获奖情况等。

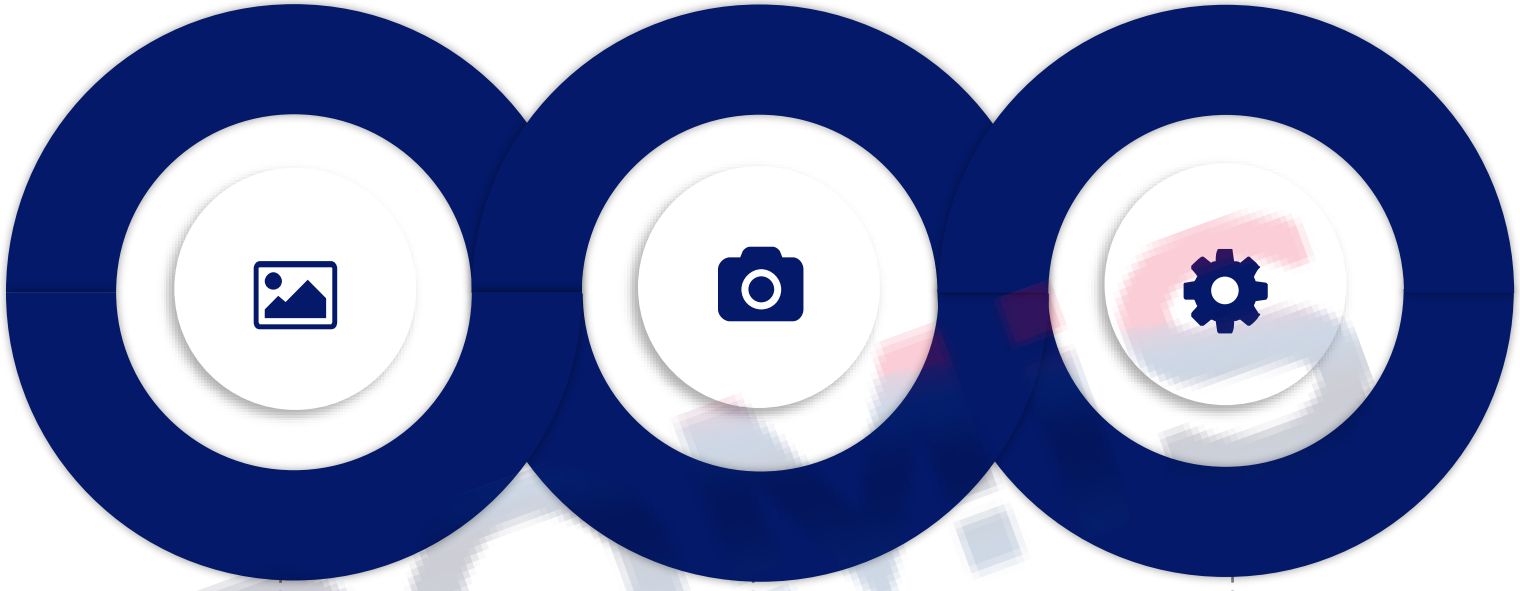


更细一点，兄弟院校近几年获得了那些重大或重点项目，等等。

深度分析和挖掘排行榜表面**数字**所反映的内涵和问题

平台建设背景

DQMiS



社会上的各类**大学（或学科）排名**，尽管我们不看重，但它的社会影响较大，特别是**影响一大批利益攸关方的选择和决定**，（北大各学科也存在、人文社科投入有限）包括学生择校、人才引进、社会资金流向等

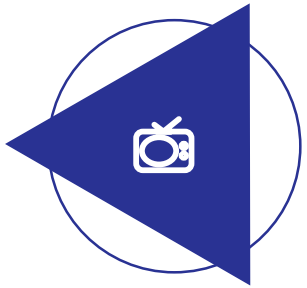
教育部颁布的《国家教育规划纲要》已经明确提出：鼓励专门机构和社会中介机构对高校学科、专业、课程等水平和质量进行评估。**建立科学、规范的评估制度。**探索与国际高水平教育评价机构合作,形成中国特色大学评价模式

2017年12月，教育部学位与研究生教育发展中心发布了**第四轮**全国性的一级学科评估结果

02

平台建设的基本内容

平台建设目的



为**学校发展和学科建设**提供全方位的咨询服务，具体包括：学科定位、学科发展重点、学校（学科、院系）各项基本指标的动态监测等。

纵向定量地分析北京大学各学科（或院系）近十年中各基本指标的发展状况。

横向定量展示各学科在国内外的位置。与兄弟院校相比，客观、准确地诊断自身存在的优势和不足。定量地指明我们的哪些指标占据优势，哪些指标处于劣势地位，距离是多少？进一步明晰努力的方向。

平台建设的主要内容

DQMiS

就北大而言，分类别构建不同类学科的评估**指标体系**。

提供构建事实型数据库管理系统、统计分析、结果综合展示的实施方案，并进行系统开发。可提供校内数据分析的在线服务。**动态监测学校**各项基本指标的发展变化情况。

研究对大学、学科（或院系）进行**诊断、分析、评价的理论与方法**。

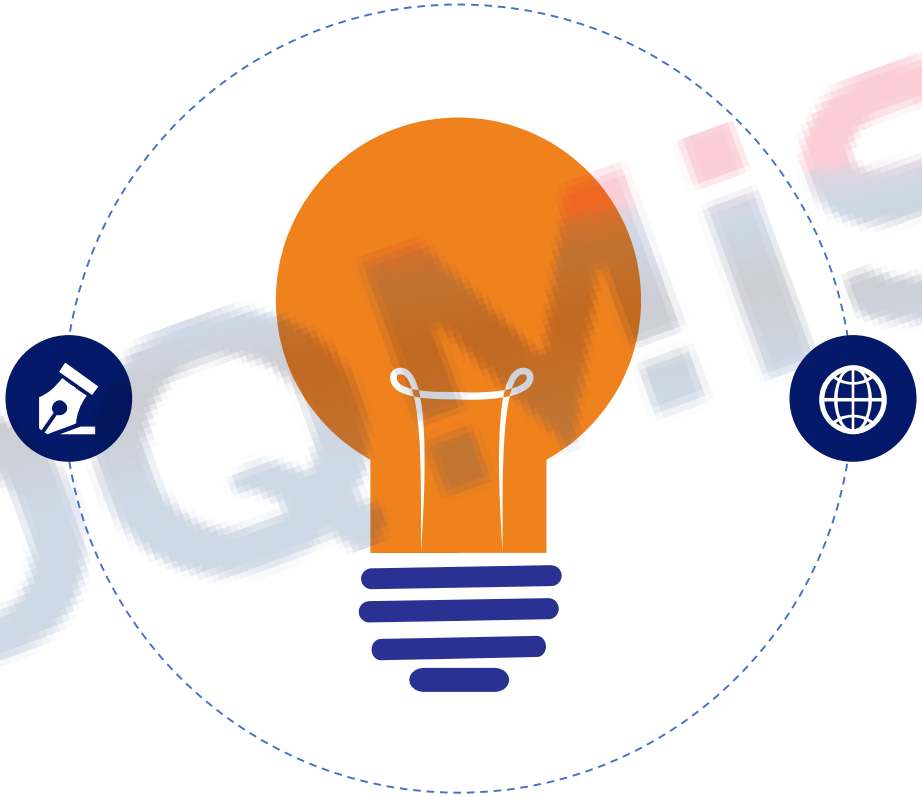
研究如何**获取各类指标数据**，这包括北大及与北大有竞争关系的兄弟院校。如何对这些数据进行有效的集成与处理。这是本课题的工作重点与难点之一。（搜集+购买）

研究诊断分析**结果的展现**方法及呈现的形式。如有必要，开展基于学科内容的文献计量学分析。

学科门类与一级学科

学科门类

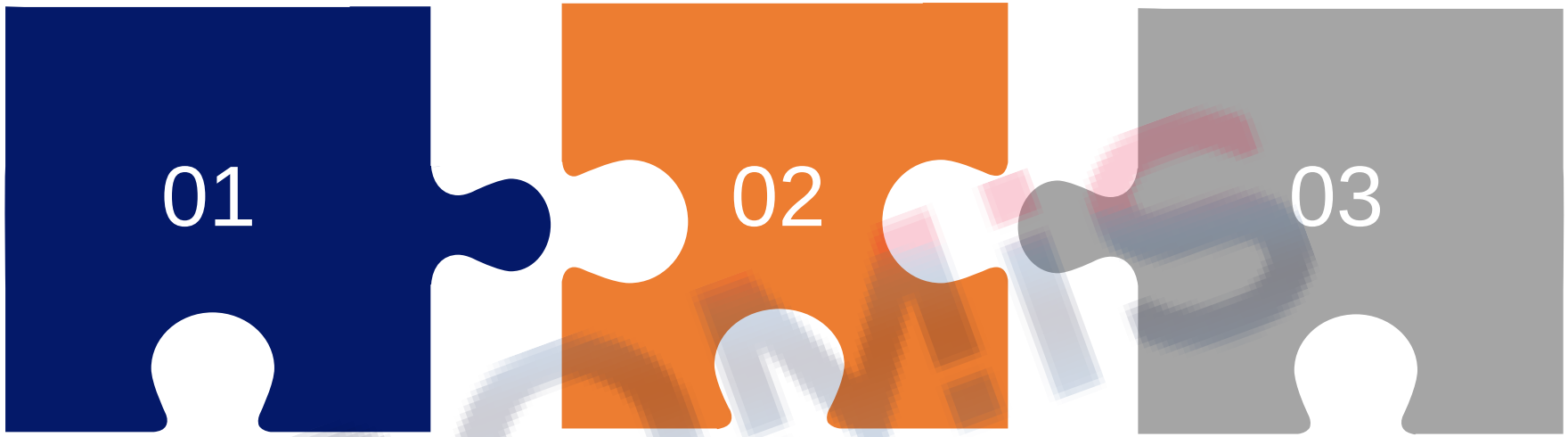
2011年3月，国务院学位委员会和教育部颁布修订的《学位授予和人才培养学科目录（2011年）》，规定我国分为：哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学、艺术等13个学科门类。110个一级学科。



“一级学科”

它是指国务院学位委员会根据科学研究对象、范式、知识体系和人才培养的需要划分的学科分类体系。

理工科类、人文社科类的一级学科



理科类的一级学科

数学、物理学、化学、天文学、地理学、大气科学、地球物理学、地质学、生物学、生态学、统计学等。

工科类的一级学科

力学、环境科学与工程、核科学与技术、电子科学与技术、信息与通讯工程、计算机科学与技术、软件工程等。

人文社科类一级学科

中国语言文学、外国语言文学、哲学、社会学、理论经济学、应用经济学、政治学、图书情报与档案管理、马克思主义理论、新闻传播学、法学、教育学、体育学、中国史、世界史、考古学、民族学等

基本思路

DQMiS

研究对象	•人文社科类一级学科：中国语言文学、中国史、哲学、法学等 •国外相似学科：国外对应学科或相关学科，如QS大学学科分类等
指标体系	•调研现有指标：研究国内外各类排行榜以及教育部一级学科评估等各类指标 •确定指标项：根据不同的学科类型，分别确定需要收集与分析的基础指标项
数据搜集	•网络数据：教育部、科技部、国家社科、国家自科基金等科研机构网站 •文献数据库：Web of Science(SCI、SSCI&AHCI)、CSSCI、同方、万方、维普等
数据加工	•数据处理：数据清洗、变换、集成、规范、分析、映射关系模型的构建等 •平台构建：基于评估指标数据的平台开发、测试、运行、自动报表生成等
应用服务	•分析报告：高校各学科基本指标的情况、与国内兄弟院校的对比分析等 •决策分析：学科优劣势分析、竞争力分析、学科演化与发展态势分析等

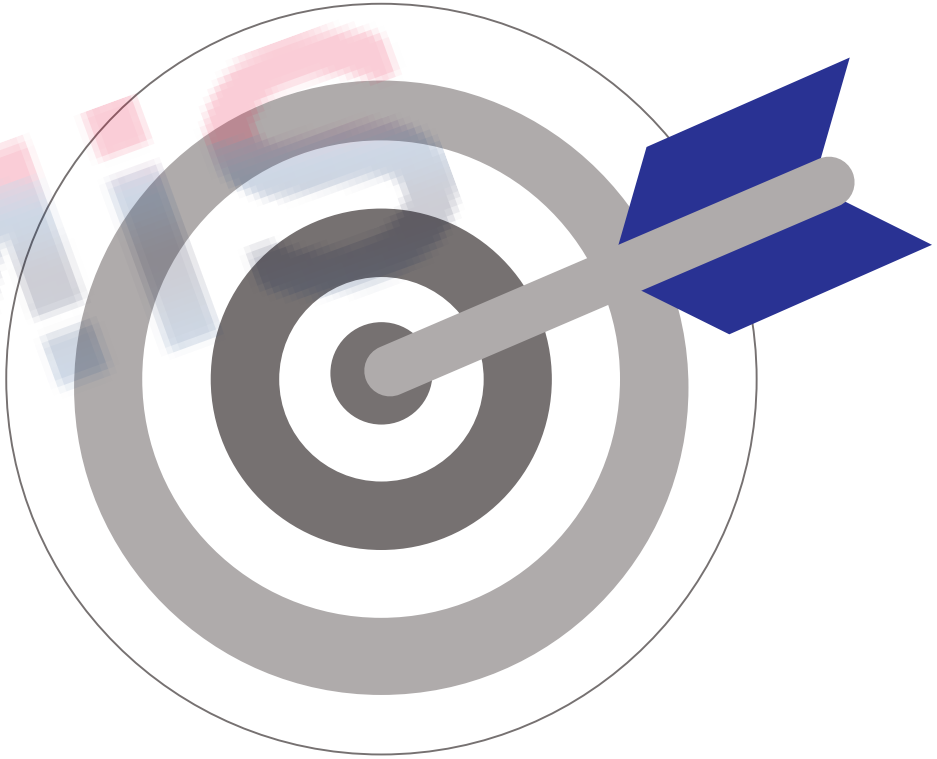
02

数据搜集与处理方法

数据采集与规范--科研项目数据（2002—2015年）

DQMiS

- **国家社科基金**数据：收集科研项目约3万余条，数据完整齐全。
- **教育部人文社会科学基金数据**：收集科研项目2万余条。数据繁杂，各种格式（pdf,excel等格式，发布日期不定）如：语言学 —— 中国语言文学？外国语言文学？
- **国家自然科学基金数据**：没有一级学科对应关系。需要结合人名和机构名称映射出相应的学科门类。（管理学 ？ 图书情报档案）



数据采集与规范--科研成果数据（2002—2015年）

国际优秀论文数据（SCI、SSCI、A&HCI）

数据处理：抽取作者、年份、标题、关键词、地址、期刊等字段，其中，将作者所在的机构进行抽取，并进行统一。

- 如何确定学科文献？国内外学科对应问题。
- peking univ和Beijing univ → 北京大学
- tsinghua univ和qinghua univ → 清华大学

- 哪些为一级学科对应的SCI 论文？

获奖数据的处理

一等奖	社会学	中国农村留守人口研究：别样童年、阡陌独舞、静寞夕阳	著作奖	社会科学文献出版社	2008 年 8 月	叶敬忠、潘璐、吴惠芳、贺聪志

拆分为

一等奖	社会学	中国农村留守人口研究：别样童年、阡陌独舞、静寞夕阳	著作奖	社会科学文献出版社	2008 年 8 月	叶敬忠
一等奖	社会学	中国农村留守人口研究：别样童年、阡陌独舞、静寞夕阳	著作奖	社会科学文献出版社	2008 年 8 月	潘璐、吴惠芳、贺聪志

映射关系模型

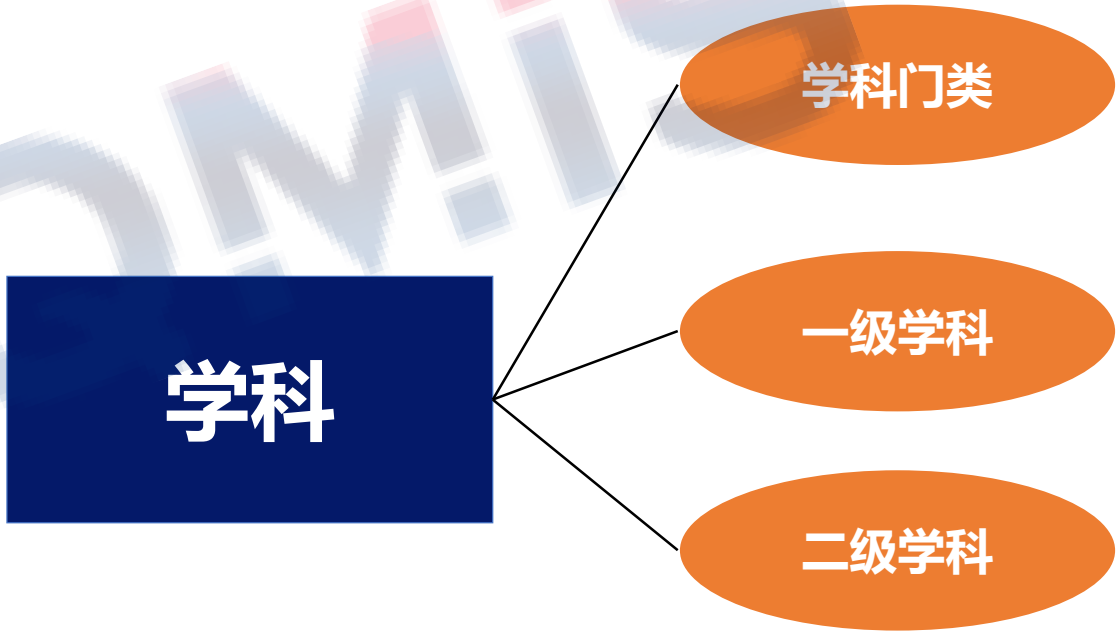
“学者—学科---机构”，如：林毅夫—理论经济学—北京大学

映射关系模型

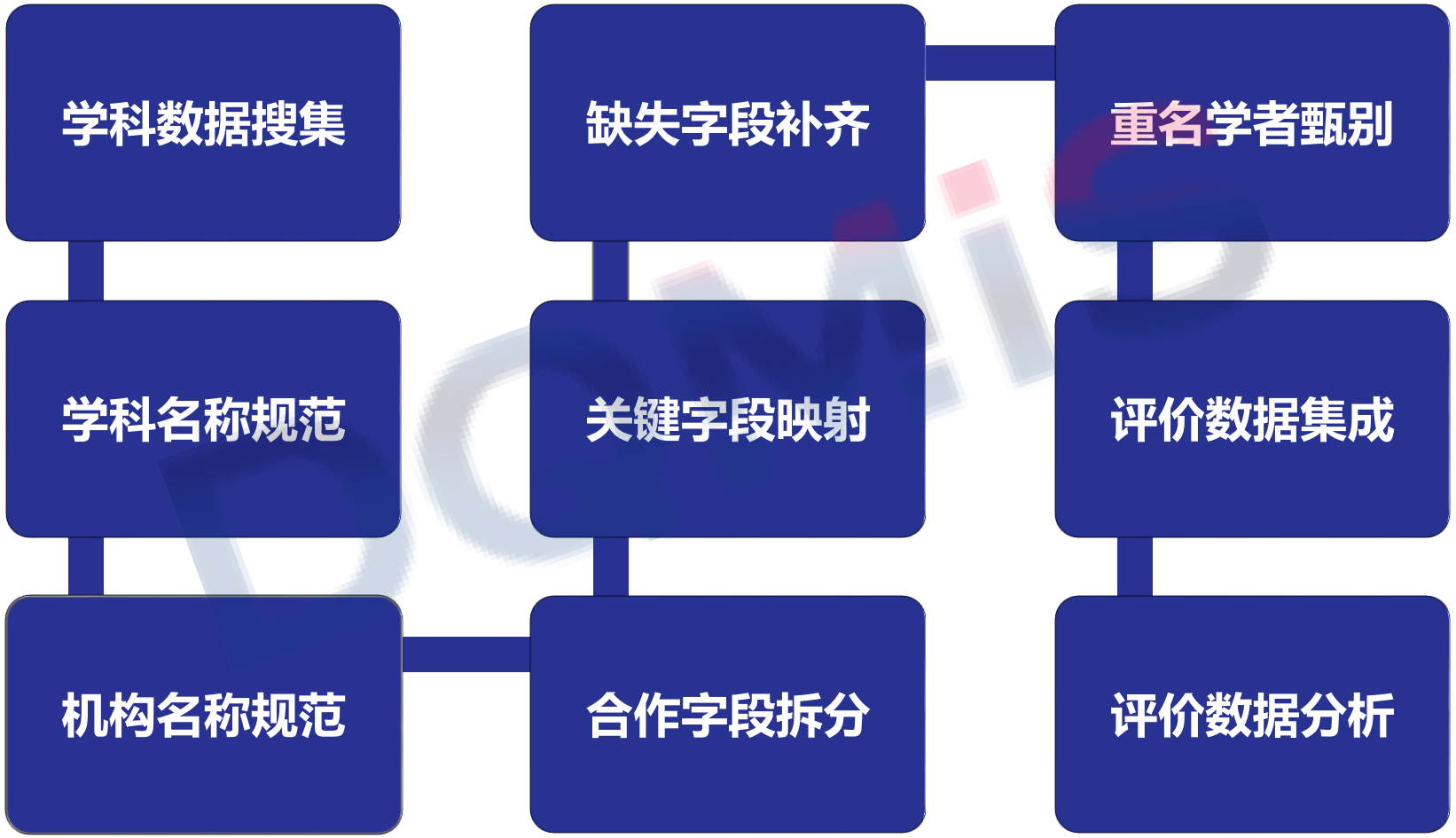
长江学者 ：

北京大学　谭文长　流体力学

北京大学　黄桂田　政治经济学



数据处理流程（9个步骤）


学科数据搜集
缺失字段补齐
重名学者甄别
学科名称规范
关键字段映射
评价数据集成
机构名称规范
合作字段拆分
评价数据分析

04

主要功能与成果

平台的主页： http://scie.pku.edu.cn/

北京大学一级学科数据综合分析平台

DQMiS

专用服务器

网址： http://scie.pku.edu.cn

系统平台的功能指标

对任意一个一级学科，系统可自动生成任何一所高校的学科数据综合分析报告，以及10余所兄弟院校主要科研指标对比分析报告；

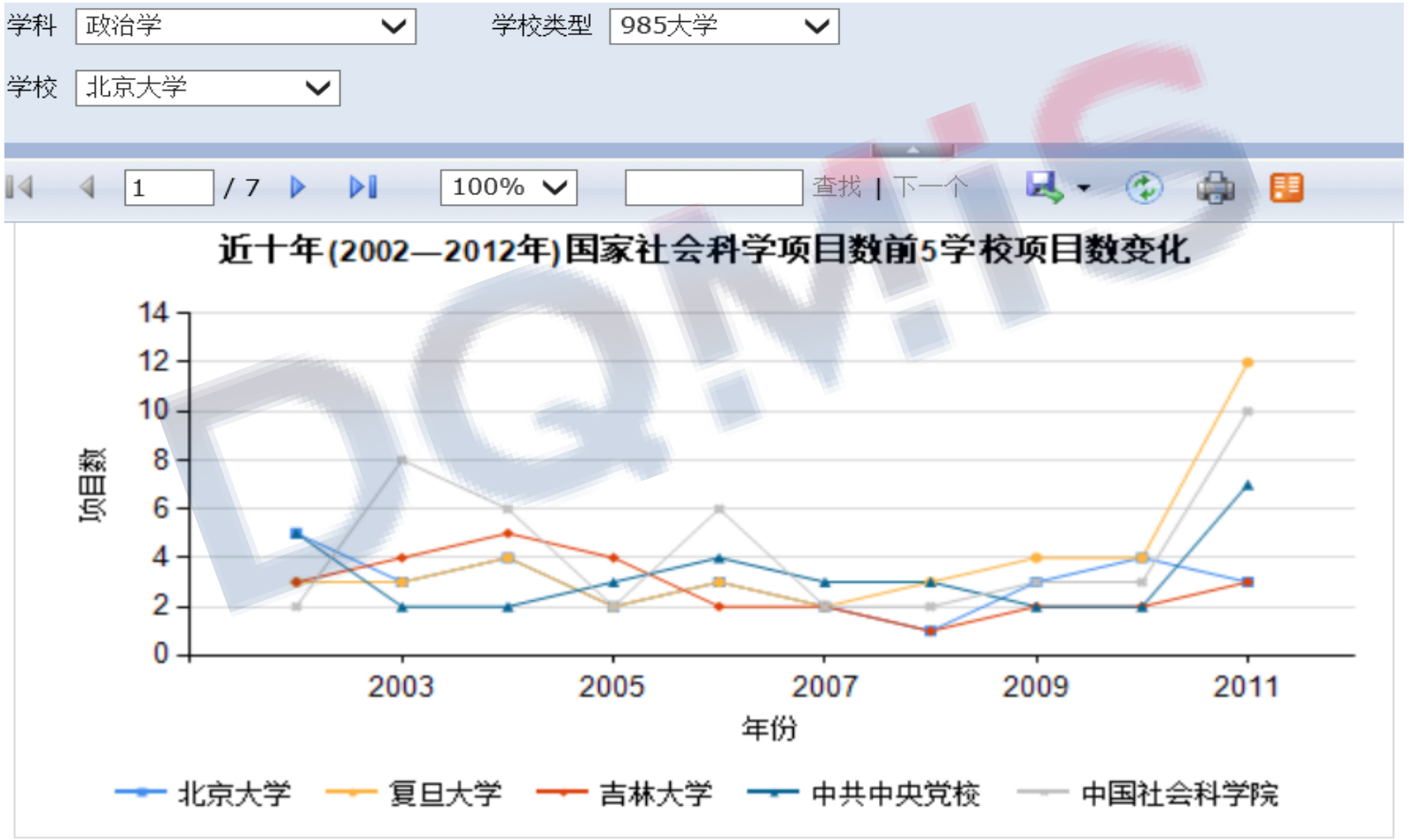
系统性能指标

任何单一学科数据报告的产生时间不超过1分钟

基本功能

DQMiS

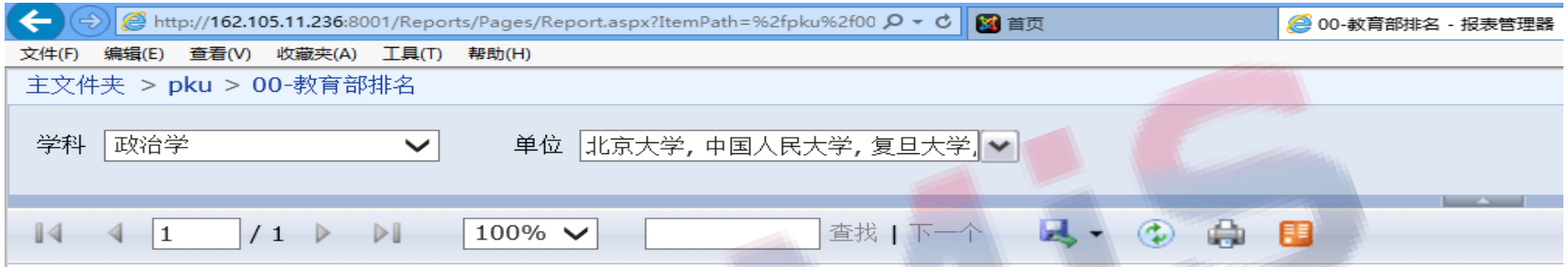
该平台可自动产生17个学科中任何一所学校的学科分析报告



基本功能

DQMiS

可产生国内若干所大学主要量化指标的对比分析一览表



教育部一级学科（政治学）数据对比分析

教育部排名	大学或可研机构	国家社会科学基金项目	国家自然科学基金项目	教育部人文社科项目	SCI、SSCI和AHCI	中文核心期刊	优秀博士论文	一级学科	二级学科	研究基地
1	北京大学	30	1	20	44	2933		1		1
1	中国人民大学	28		23	24	4281	1		2	1
3	复旦大学	40	5	27	40	1430	1		2	1
3	华中师范大学	21		28		1049	1		2	
5	南开大学	25	1	20	6	1043			1	1
5	中共中央党校	33				739	1		2	
8	南京大学	17	2	13	2	1211				
12	华东师范大学	18		14	4	744				1

制定一套标准规范

北京大学
PEKING UNIVERSITY

北京大学学科评价流程规范与数据标准

北京大学信息管理系科学评价研究组

2013 年 9 月 1 日

目　录

获得一项软件著作权、一项专利

DQMiS

北京大学一级学科数据分析平台

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第0640194号

软件名称：北京大学一级学科数据分析平台软件 V1.0

著作权人：北京大学

开发完成日期：2012年09月16日

首次发表日期：2012年10月25日

权利取得方式：原始取得

权利范围：全部权利

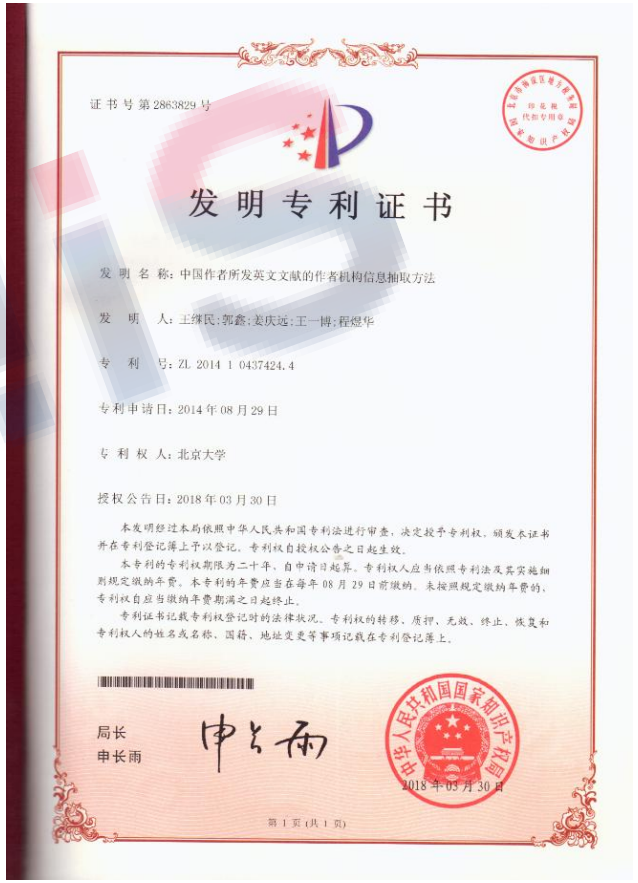
登记号：2013SR134432

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

No. 00365030

2013年11月28日



证书号第2863829号

发明专利证书

发明名称：中国作者所发英文文献的作者机构信息抽取方法

发明人：王继民;郭鑫;姜庆远;王一博;程煜华

专利号：ZL 2014 1 0437424.4

专利申请日：2014年08月29日

专利权人：北京大学

授权公告日：2018年03月30日

本发明经过本局依照中华人民共和国专利法进行审查，决定授予专利权，颁发本证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。

本专利的专利权期限为二十年，自申请日起算。专利权人应当依照专利法及其实施细则规定缴纳年费。本专利的年费应当在每年08月29日前缴纳。未按照规定缴纳年费的，专利权自应当缴纳年费期满之日起终止。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长 申长雨

中华人民共和国国家知识产权局

2018年03月30日

第1页(共1页)

开展的学科应用服务

- 为国内10几所高校提供学科数据分析服务
- 主要有：北京大学、浙江大学、华东师范大学等

学术论文、著作等

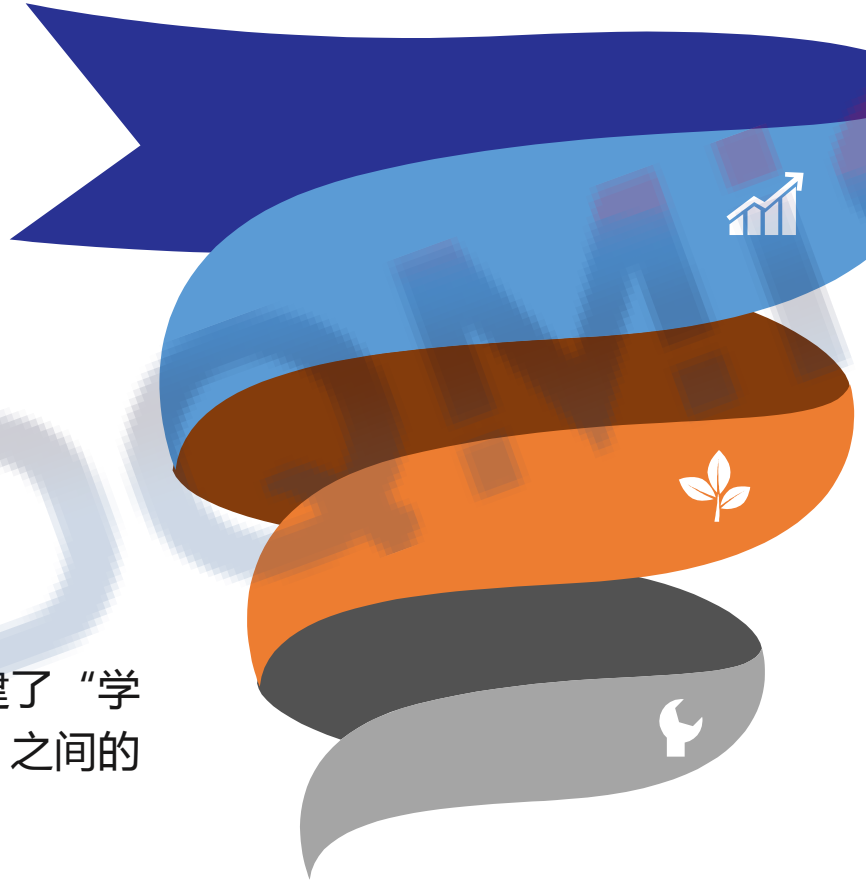
王继民，刘洋，徐怡. 高校一级学科评估数据的集成方法与实践. 第十三届全国科技评价学术研讨会. 上海. 2013年11月.

图书一部：中国人文社科类一级学科数据分析报告. 科学出版社. 2014年

特点

DQMiS

提出了一套适合我国人文社科类一级学科评估数据的收集、处理、集成、规范，以及数据自动更新的方法与技术；



构建了一个公开、透明、集成了人文社科类一级学科评估数据的综合分析平台，可为我国高校的学科建设提供基本的数据支撑。

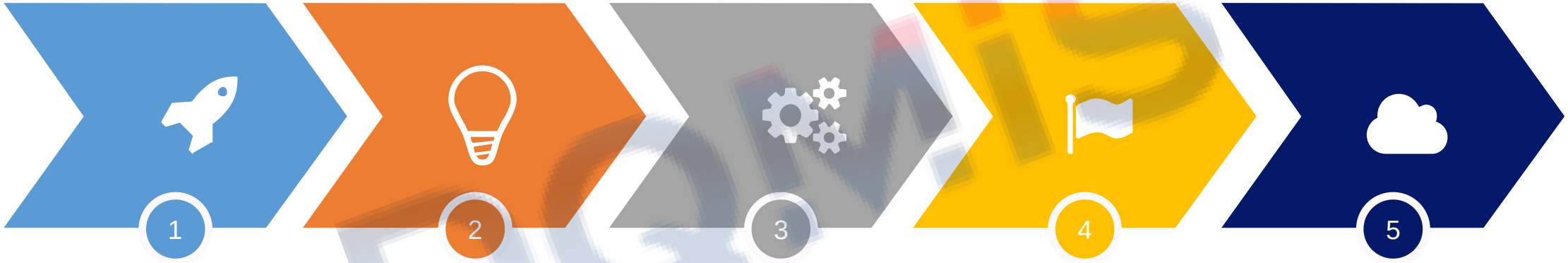
分学科门类，构建了“学者—学科—机构”之间的映射关系模型；

05

数据驱动的评价方法与实践

数据驱动的评价方法与实践

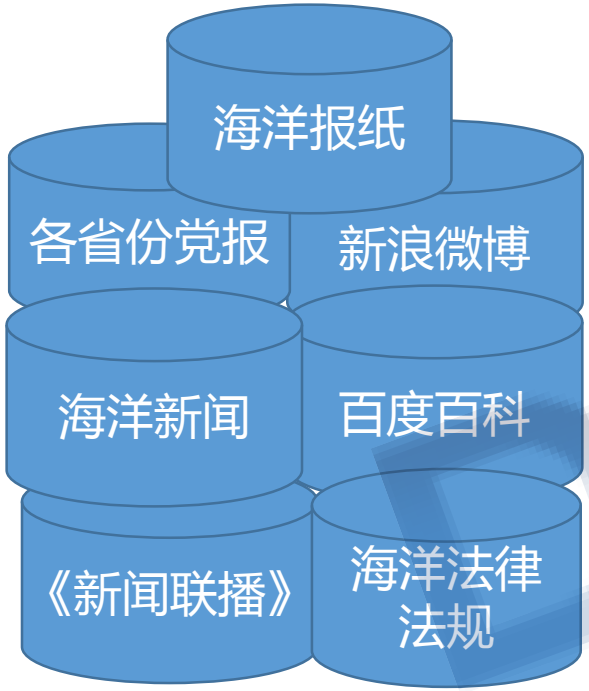
提出了一套以多源数据为基础的、以海洋意识为主题的、基于数据驱动构建评价指标体系的方法，并通过实证研究验证了该方法的有效性。



1. 根据文献、电视新闻、报纸、网页、微博等多种涉海数据，构建了一个海洋主题词表和海洋词向量模型（Word2Vec），为后续中文分词和关键词的扩展奠定了基础。
2. 基于海洋意识主题词的共现关系与聚类分析结果，构建了一个海洋意识评价指标体系，并检验了该指标体系的有效性。
3. 测算了全国大陆地区31个省份的海洋意识的总得分与四个一级指标的得分，所得结果与2016年的排名结果具有很强的相关性。
4. 测算了全国各省份的333个地级市的海洋意识得分，这是传统方法难以得到的结果。
5. 建立了一个海洋意识数据可视化平台，实现了对31个省份、333个地级市海洋意识得分的对比分析和可视化展示。

海洋词表和海洋词向量模型

数据源



《新闻联播》、报纸、新闻等数据源的数据总量约为31万条

数据采集和处理

- 机器处理：
- 抓取工具+程序提取
- 数据预处理
- 数据集成
- 数据标准化

海洋词表

海洋辐射传递
海洋腐蚀
海洋腐殖质
海洋港口
海洋高等教育
海洋高技术
海洋高技术产业竞争力
海洋高能环境
海洋高新技术
海洋高新技术产业

集成知网文献关键词、国家标准等数据源，得到6575个词的海洋词表。

海洋词表和海洋词向量模型

DQMiS

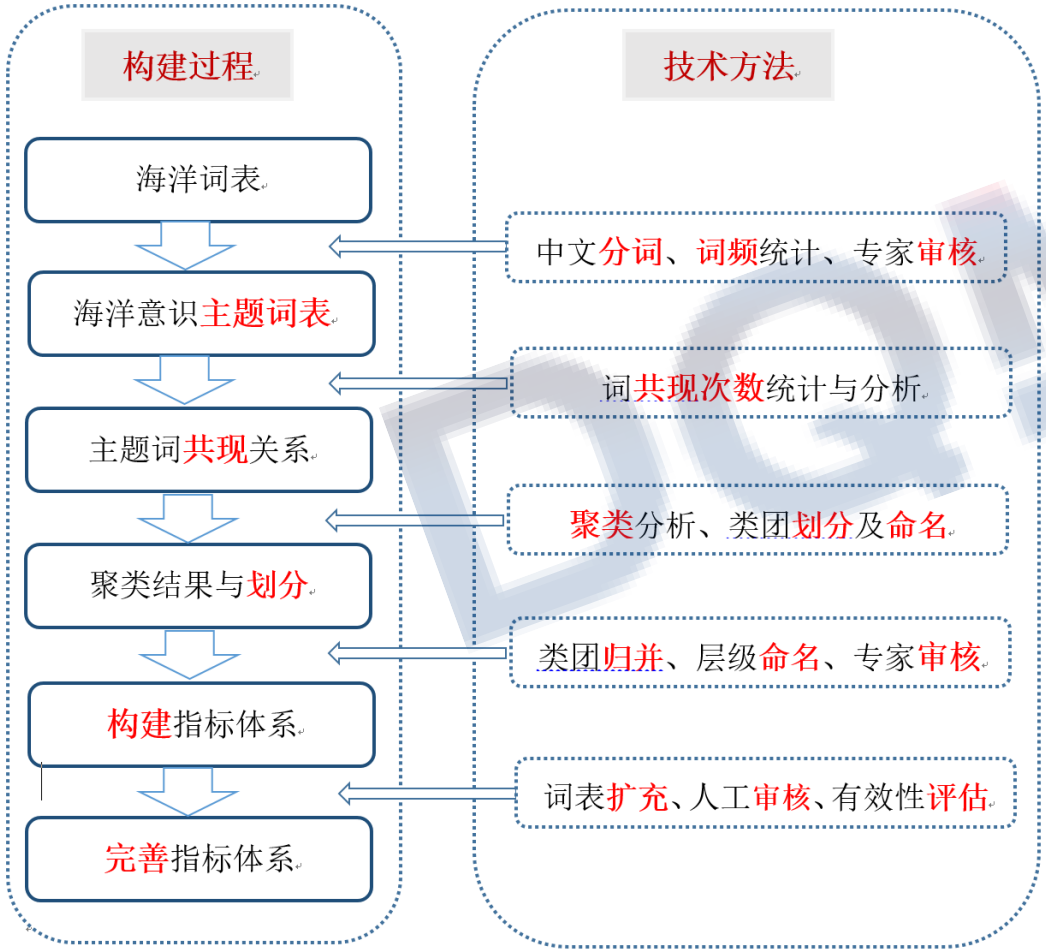
海洋词向量模型

- 构建了一个Word2Vec词向量模型
- 利用模型可对**涉海词**进行扩展

数据驱动的海洋意识指标体系的构建

DQMiS

指标体系的构建过程



指标体系的构建方法

- ✓ **共词**分析方法
- ✓ **聚类**分析方法
- ✓ **社会网络**分析方法

指标体系的有效性评估

- ✓ **效度系数法**

实证研究

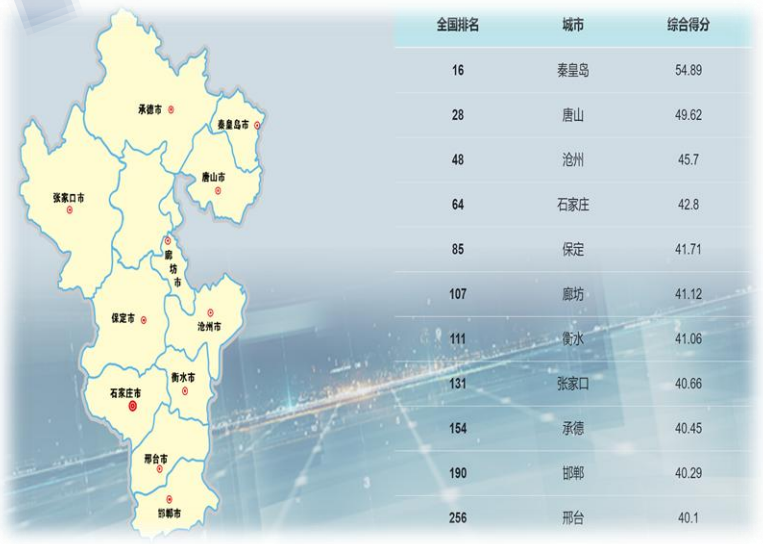
DQMiS

研究对象

- 全国**31个省份**海洋意识的一级指标得分、综合得分与排名。
- 全国**333个地级市**的综合得分及排名

计算过程

- 先计算三级指标得分，之后**自底向上**归约，得到综合得分。



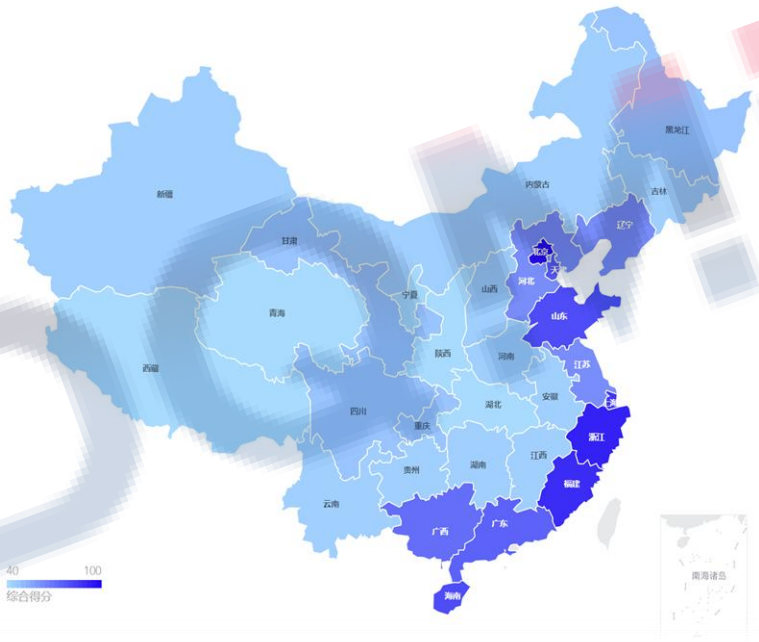
全国排名	城市	综合得分
16	秦皇岛	54.89
28	唐山	49.62
48	沧州	45.7
64	石家庄	42.8
85	保定	41.71
107	廊坊	41.12
111	衡水	41.06
131	张家口	40.66
154	承德	40.45
190	邯郸	40.29
256	邢台	40.1

实证研究---各省份综合得分测算

DQMiS

部分结果

排名	省份	综合得分
1	北京	94.83
2	浙江	86.41
3	福建	86.25
4	上海	82.34
5	海南	79.02
6	天津	73.60
7	山东	73.42
8	广东	69.83
9	广西	65.86
10	江苏	60.51
11	辽宁	59.15
12	河北	57.62
13	甘肃	47.15
14	重庆	46.94
15	四川	46.74
16	黑龙江	45.73



结论

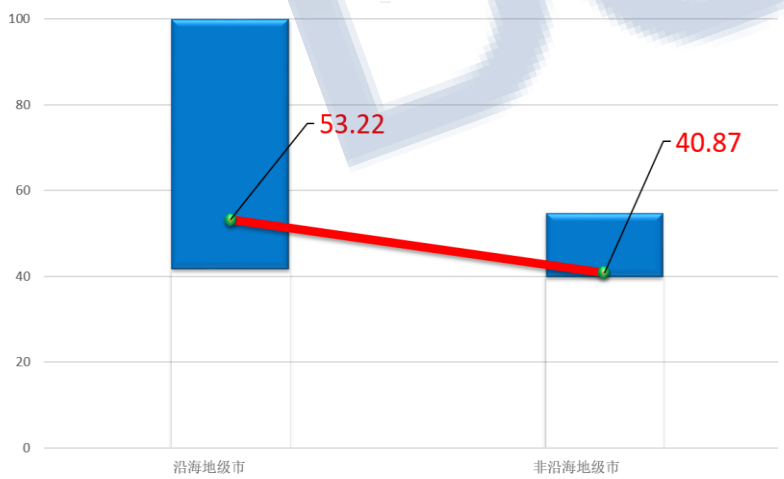
- 整体来看，各省份海洋意识的发展水平存在**较大差异。**
- 使用相关系数来衡量所计算结果之间的关系。通过计算Pearson相关系数的结果为0.76 ，表明本文的测算结果与2016年《国民海洋意识发展指数》中的测算结果呈强相关。
- 我国31个省份中，海洋意识综合得分最高的5个省份分别是北京、浙江、福建、上海和海南。

基于海洋意识指标体系的实证研究---各地级市的综合得分测算

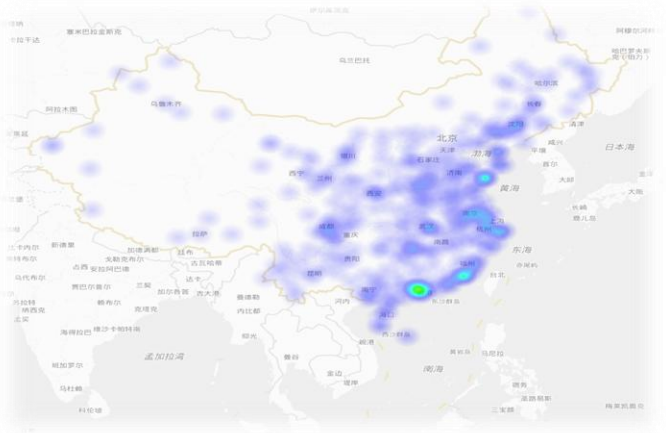
DQMiS

结论

排名	地级市	综合得分
1	青岛	100.00
2	厦门	90.47
3	大连	77.75
4	广州	73.60
5	舟山	66.65
6	深圳	66.07
7	宁波	65.18
8	三亚	62.21
9	海口	60.85
10	福州	57.31



- 整体来看，各地级市的海洋意识的发展水平存在**较大差异**。
- 我国333个地级市中，海洋意识综合得分最高的5个地级市分别是青岛、厦门、大连、广州和舟山。
  - ✓ **沿海地级市**的海洋意识发展水平高于非沿海地级市。
  - ✓ **东部地级市**的海洋意识发展水平高于中西部地级市。



数据可视化平台的构建

DQMiS

所用技术

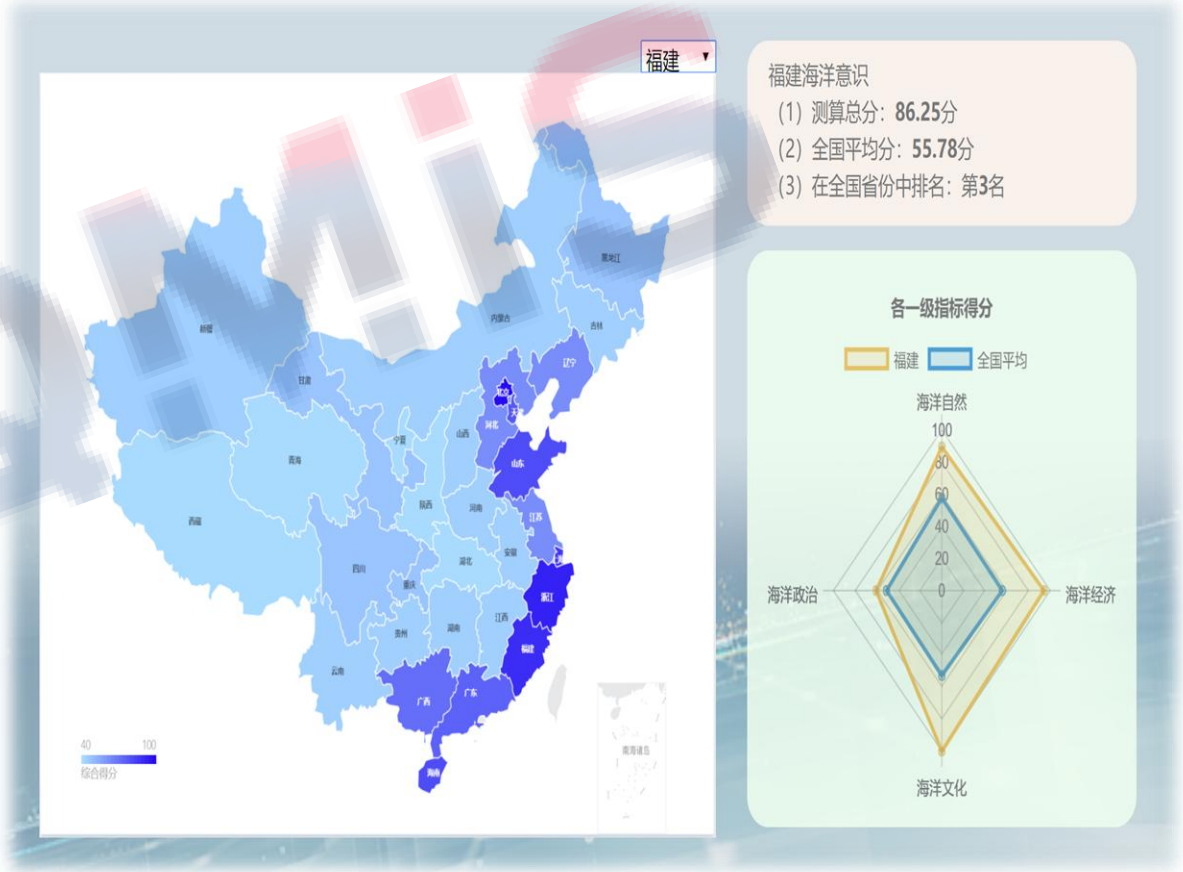
Python 中Flask Web框架

平台内容

对于用户输入的省份，

- ✓ 各省份海洋意识综合得分与排名情况。
- ✓ 海洋意识全国平均分。
- ✓ 各省份海洋意识一级指标得分。

平台展示



DQMiS

第二届数据质量管理国际峰会

The 2nd Data Quality Management International summit

北京大学 PEKING UNIVERSITY

Thank You!