

“链科创” 数据库手册

同济大学上海国际知识产权学院

创新与竞争研究中心

1. 数据库结构

当前，以科研人员为核心的研究议题正在不断丰富，以科研人员为核心的数据库产品正在蓬勃发展。与“中国知网”“ResearchGate”“Linkedin”“科研之友”“迈科技”等国内外大型学术/商用数据库不同，“链科创”数据库（以下简称数据库）聚焦“高校教师数据画像”，多源匹配教师简历信息、科研项目数据、学术论文数据、专利申请数据、师生网络数据、衍生企业数据，形成以人为核心的高校教师创新创业数据库。“链科创”数据库的最大特色是数据整合与多维展现，能够回答“科研人员是谁？科研人员做了什么？”两个问题。

“链科创”数据库的结构示意图如图 1 所示。

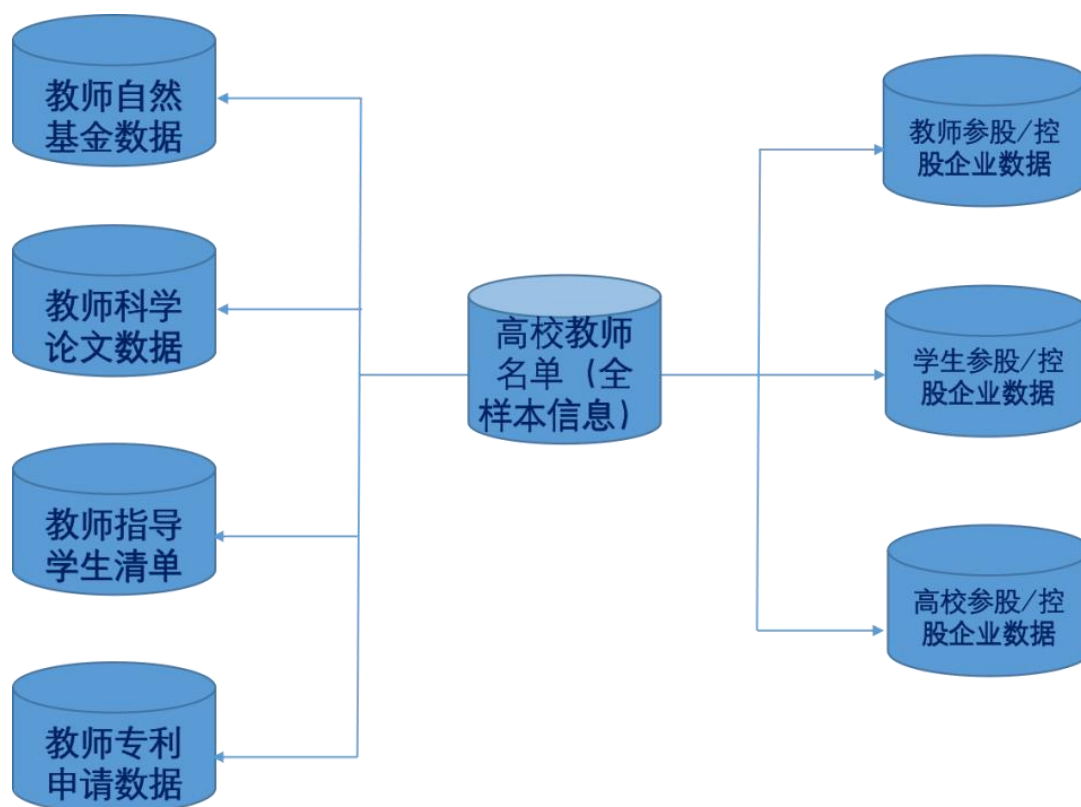


图 1 “链科创”数据库的数据构成

截止 2020 年底，“链科创”数据库各组成部分的数据规模如下：

双一流高校	高校教师	硕士/博士生	国家自然科学基金
40	18.86 万	140.47 万	19.98 万
学术论文	中国发明专利	教师创业企业	高校货币出资企业
212.71 万	54 万	9 万	1.7 万

2. 数据库申请与使用规则

“链科创”数据库致力于打造公共科研数据平台，服务对象包括政府部门、高校院所科研管理部门、企事业单位、科研人员个人等。

数据库申请与使用规则如下：

（1）非学术用户可以直接在数据库网站 www.linkfaculty.com 检索查询所需相关信息，检索词列表详见附录；

（2）学术用户可以直接联系我们，描述数据需求，数据库团队会根据需求提供特定格式的数据；

（3）学术用户使用数据库提供的数据发表学术论文，请将数据库团队对接人员列入作者名单，或者在文末特别予以说明数据来源；

（4）数据库所有数据属于同济大学，未经数据库团队授权，所有用户不得利用网络爬虫等工具下载、传播数据库数据。

3. 数据库共享共建与数据标准

我们欢迎用户一起共同搭建以人为核心的科研人员数据画像平台。任何向“链科创”数据库贡献有效数据的组织或个人都将自动成为数据库成员，并获得数据库底层数据的使用权限。

数据库共享共建的文件格式包括 excel、csv 等。

数据采集标准如下：

3.1 高校教师个体数据

高校教师的个体数据主要来自于高校下属二级学院官方网站，包括教师所在学校名称、学院名称、性别、年龄（如果有）、职称、职称晋升的时间点、邮箱信息。针对样本高校内部因职位变动（如从 A 学院转岗到 B 学院）、双聘岗位（如同时担任 A 学院和 B 研究中心教师）的情况，为减少识别误差，避免重名干扰，建议不采集高校内部冠以“研究中心”“研究院”等非传统学院内的教师信息。

教师信息表如下：

教师编号	教师姓名	学校	学院	性别	年龄	职称	主页链接	晋升正高职称时间	晋升副高职称时间	邮箱	是否明星企业家

3.2 专利数据

基于中国国家知识产权局的专利数据库，以高校“教师姓名为第一发明人”作为检索条件，获得教师全部的专利申请信息。

专利信息表如下：

专利名称	公布号	申请人	地址	发明人	分类号	申请日	摘要	法律状态

3.3 国家自然科学基金数据

采集高校所有年份的国家自然科学基金数据，数据字段包括基金名称、承担单位、承担人、批准号、所处领域等信息。

自然科学基金项目信息表如下：

批准号	项目名称	依托单位	题名	申请领域	基金类型	资助金额	立项时间	是否结项

3.4 学术论文数据

采集了教师基于国家自然科学基金资助完成的学术论文数据，包括作者、作者单位、国家自然科学基金批准号、合作者、合作单位、通讯作者邮箱、论文题目、发表年份、总被引用率、期刊检索、期刊影响因子、期刊分区等论文信息。

学术论文信息表如下：

作者	作者单位	通讯作者	通讯作者单位	邮箱	论文题目	发表年份	文献来源	基金资助号	影响因子	总被引频次	DOI	JCR分区

3.5 硕士生/博士生数据

通过中国知网（CNKI）博硕士学位论文库，以高校名称作为单位检索词，下载高校在库的硕士/博士论文信息，包括论文题目、学生姓名、指导教师姓名、授予学位单位、授予学位时间信息，从中提炼出“教师—学生”关系数据。

需说明的是，由于部分高校存在某些年份硕士/博士论文不上传中国知网的情况，由此造成一定程度的数据缺失。

硕士/博士信息表如下：

教师姓名	高校名称	硕士/博士	学生姓名	论文题目	毕业年份

3.6 教师创办衍生企业数据

以高校为地理中心点，基于企查查数据库，提取以中心点 1 公里范围内的所有企业数据，包括企业成立年限、注册资本、法人代表姓名、股东姓名、董监事姓名、行业领域、主营业务范围等信息。基于采集的企业法人代表、股东、董监事信息，与教师姓名进行匹配，获取教师的创业情况。

需重点说明的是，即便限缩在 1 公里范围内，教师姓名与企业法人代表、股东、董监事姓名也依然存在重名问题，因此，姓名匹配工作最大的难点是“如何克服重名问题”。我们的做法如下：（1）高校端删除样本高校内部所有重名的导师，如“某强”“某建军”等极易重名的姓名，以确保参与匹配教师姓名的唯一性，消除教师-企业

人员姓名匹配成功后的识别歧视；（2）企业端限定行业类型“科学研究和技术服务业”“信息传输、软件和信息技术服务业”“教育”“金融”“文化、体育和娱乐业”“制造业”六个行业，排除其他行业企业相关人员姓名对匹配工作的干扰。

教师创办企业信息表如下：

教师姓名	企业名称	法人/股东	成立时间	注册资本	所属行业

3.7 硕士/博士创办衍生企业数据

本研究聚焦样本高校硕士/博士研究生毕业后留在高校周边创办衍生企业的情况。相应的数据匹配同教师创办衍生企业的匹配方法。

硕士/博士创办企业信息表如下：

硕士/博士姓名	企业名称	法人/股东	成立时间	注册资本	所属行业

3.8 高校货币出资衍生企业数据

本研究基于企查查数据库，以样本高校所对应资产经营公司的名称，通过股权穿透，获取了高校货币出资的所有衍生企业名单，包括衍生企业基本工商信息、样本高校持股比例、衍生企业法人代表、股东名单、董监事信息等。

高校货币出资企业信息表如下：

高校名称	企业名称	法人/股东	成立时间	注册资本	所属行业