



# 中国开放数据 使用手册

2021版

# CnOpenData 团队简介

## Our Team



Hello! 我们是一群IT男，“技术宅”是大家给我们的标签，但我们喜欢这种没事写写代码、抓抓数据的欢乐时光。

一次偶然的机会让我们和一些高校老师建立了联系，为他们提供数据搜集和整理服务。过去几年，我们和不少高校老师建立了稳定、愉快的合作关系，在小范围内竟然产生了一定的影响力，越来越多的学者开始联系我们希望提供数据服务。随着客户数量的增加，我们萌生了建立数据平台的想法。

互联网产生了海量数据，记录着丰富信息，并以开放的姿态呈现出来，然而由于数据量巨大、搜集难度极高、清理整合耗时等问题，学者在使用这些数据时面临诸多障碍，而这恰是我们这些IT男最为擅长的地方。当然，我们也有强大的学术团队做支撑，他们曾经是我们的客户，今天成为我们的顾问！

简约、清晰、有用，是我们追求的理念。

数据是散落的金子，我们愿做挖矿者，期待与作为炼金人的您精诚合作！





## 1. 专利创新数据

- 提供1985年-2020年的专利申请和专利授权数据，及专利事务数据，分发明公布、发明授权、实用新型、外观设计等，详细追溯专利的各个事务状态；
- 含主表、事务表；
- **字段**：申请公布日、PTC公布及申请数据、申请人、申请日等。

## 2. 中国专利引用数据（中国版、世界版）

- 提供1985年-2019年所有中国专利在国内外的引用与被引用详细情况（世界版现已更新至2020年，中国版正在更新）。世界版包含中国专利被本国申请的专利引用情况、世界上其他国家对中国专利的引用情况、该专利被引用情况的详细信息、该专利引用其他专利的详细信息以及各个专利在每个国家截止目前的法律状态信息；
- 含基本信息表、引用表、被引用表、事务表；
- **字段**：专利ID、IPC分类号、引用及被引用专利优先权日、事件类型等。

## 3. 绿色专利创新数据

- 根据废弃物管理类、核电类、交通运输类、能源节约类、农林类、替代能源生产类、行政监管与设计类7大类别分别分表制作，各自形成发明公布、发明授权、实用新型三类，结构清晰；
- 含发明公布表、发明授权表、实用新型表；
- **字段**：申请号、网站标题、PTC申请及公布数据、本国优先权等、授权公布号等。

## 4. 中国专利文本数据

- 数据总量超2300万的大样本量文本类数据，时间区间为1985-2019年，主体为专利描述文本信息和专利权利要求文本信息；
- **字段**：专利ID、描述内容。

## 5. 中国工业企业、税收调查企业专利及引用被引用数据

- 参考科学的数据匹配与处理方法，将中国工业企业/税收调查企业数据与中国专利创新数据进行细致匹配而成。中国工业企业数据截至2013年，专利数据截至2018年，税收调查企业数据时间区间为2011年-2015年；
- 发明公布/发明授权/实用新型/外观设计四类表下详细分为：基础信息表、引用表、被引用表、事务表；
- **字段**：公司名称、年度、申请公布日、专利代理机构、被引用专利申请公布号、申请号等。

## 6. 中国工商注册企业全信息

- 极具研究价值且覆盖面广泛的中国工商注册企业全信息，涵盖企业的基本信息、重点关注、知识产权、企业发展、经营状况五个方面，合计超过40张表格设置；
- **字段**：控股企业、变更记录、企业年报、网站备案、专利信息、股权冻结、融资信息、行政许可等。

## 7. PPP项目数据

- 政企合作型城市基础设施项目数据，覆盖项目基本信息、准备阶段、采购阶段、执行阶段全数据；
- 各阶段有部分表格为材料文件内容，以文件名为主要显示方式。

## 8. 上市公司及子公司名称数据

- 按年度对自1990年起的上市公司与其子公司、关联公司的关系与公司名称进行了整理，发布历年上市公司及子公司名称数据；
- **字段**：证券代码、会计年度、关联公司的名称、关联公司与上市公司的关系等。

## 9. 证券交易所问询函数据-open

- 涵盖上交所的问询函件、深交所的问询函件及回复，包含2020.01.01之前的所有数据，为案例分析、大样本实证研究提供了数据基础；
- **字段**：公司代码、公司简称、发函日期、监管问询类型、标题、文档名等。

## 10. 上市公司（绿色）专利及引用被引用数据

- 提供上市公司截止到2018年底的历年专利引用及被引用详情数据/绿色专利方面研究数据；
- 发明公布/发明授权/实用新型/外观设计（绿色专利无）四部分下各设基础信息、引用及被引用、事务表；
- **字段**：专利ID、本国优先权、简要说明、专利图片、发明人、事务数据公告日等。

## 11. 上市公司年报/招股说明书/社会责任报告数据

- 近三千家上市公司企业年报/招股说明书/社会责任报告PDF文件，时间自1998年至2019年（持续更新）

## 12. 中国小区及二手房交易数据

- 涵盖我国28个省、直辖市与对应86个地级市城市住房小区分布和二手房交易信息；
- 含小区信息、二手房已成交数据、二手房在售数据三张子表；
- **字段**：省份、城市、开发商、建筑类型、房屋户型、建筑结构、房屋年限等；

## 13. 土地交易数据

- 涵盖土地出租、抵押、转让等全角度数据；
- 含供地结果基本信息表、分期支付约定信息表、土地出租表、土地抵押表、土地转让表；
- **字段**：行政区、电子监管号、支付期号、出租期限、土地抵押权人名称、原及现土地使用权人等。

## 14. 裁判文书数据

- 提供了自1996年起至2016年底的各类型案件文书，包含全方位的字段文本内容，是适用于大数据文本分析的优质文本数据源；
- **字段**：文书全文、案件类型、案号、案由、审理法院、法律依据等。

## 15. 政府采购数据

- 目前关于政府采购的相关研究仍较少，由此CnOpenData数据团队推出政府采购数据，涵盖多维度信息，为相关研究提供优质的数据资源；
- **字段：**合同金额、所属地域、供应方、采购方、所属行业等。

## 16. 人口迁徙大数据

- 包含各地区迁入来源地信息、各地区迁出目的地信息和各城市内部出行强度信息三个子模块，涵盖迁入和迁出城市、省份的总趋势和迁入迁出明细数据，为相关研究提供了优质的大数据样本；
- **字段：**迁入/出城市、迁徙规模指数、迁入/出人口占比、城内出行强度等。

## 17. 国家标准数据

- 本数据库涵盖了所有中国现行的、即将实施的和已经废除的推荐性、强制性标准；
- 含强制性标准数据表、推荐性标准数据表；
- **字段：**标准号、标准类型、标准状态、执行及主管单位、相近标准等。

## 18. 国家基金项目数据

- 包含2019年度及以前的所有国家自然科学基金和社会科学基金项目，除立项项目的相关信息外，还涵盖详细项目负责人的信息、单位情况、成果情况等多方面的内容；
- 社科数据包含2019.07.15前所有数据，自科数据包含1987.01-2020.01期间所有数据；
- **字段：**项目负责人、单位类别、结项时间、获奖情况、是否结题、6/4/2位申请代码、对应学科及学部等。

## 19. 医疗信息大数据

- 包含医院信息、科室信息和医生信息三张子表；
- **字段：**医院的详细位置/医院等级/科室的名称/所属医院/医生人数/医生个人简历/主治领域/医师级别等信息，为相关研究提供了优质大数据样本。

## 20. 中国行政区划数据

- 分为两张子表：历年县以上的行政区划信息（1981-2019）、历年县以下的行政区划变更信息（2014-2019）；
- 前者涵盖年份、地区编码、省份、地级市、县区和县区级别等字段；后者中除了上述基本信息外，还包括了变更前后的乡镇名称和编码、变更原因、批准文件等字段。

## 21. 各地区最低工资数据

- 各地区最低工资数据数据库，包含各区县最低工资划分标准（1999.06.01-2019.01.01）、金额（1999.07.01-2019.01.01）和调整日期；
- **字段：**省份、调整日期、划分标准、全/非全日制最低工资。

## 22. 新冠病毒防疫专利信息数据-open

- 涵盖新冠肺炎治疗用药、预防用药、检测和诊断试剂盒、医疗器械、防护产品、医用消毒、医疗废弃物处理、废水处理等多个领域；
- 中国防疫专利信息表/法律状态表/世界防疫专利信息表；
- **字段：**专利申请日、分类号、法律状态、申请号等。

## 23. 新冠病毒防疫标准信息数据-open

- 收录了全球范围内与新型冠状病毒肺炎防治相关的医用防护产品、传染病检验检测、医用基础设施等标准，为抗击疫情的研发、治疗、防护等工作提供标准信息支撑；
- 医用防护产品全表/医用基础设施全表/口岸卫生检疫表；
- 字段：产品类别、实施日期、发布单位、设施类别、标准编号等。

## 24. 新冠病毒防疫企业捐赠数据-open

- 涉及到新冠病毒防疫的企业捐赠数据进行整理，该数据库收录了全国各公司捐款金额、捐款去向、捐款时间等信息；
- 企业捐赠表；
- 字段：捐赠人、募捐机构、疫情捐赠额、日期等。

## 25. 全国小区新冠病毒确诊病例数数据

- 该数据库共计11868行数据，可从时间与空间两个维度探寻疫情发展趋势；
- 字段：详细地址、经纬度、及其所属省、市、县区，以及确诊人数和更新时间。

## 26. 中国银行网点全集数据

- 包含中国所有银行网点的分布信息，以银行网点持有的金融许可证状态为线索，分为金融许可证持有表/机构失控表/机构退出表；
- 字段：机构所在分局/机构名称/机构所在地/发证机关/发证日期/持证状态等字段。

## 27. 中国银行分支机构违规被处罚数据

- 除涵盖2014-2018年中国银保监会官网公布的行政处罚决定书文中全部内容外，还整合了被处罚银行的经纬度，便于定位和进行空间计量分析；
- 字段：被处罚单位/法人/案由/依据/处罚决定/做出处罚的机关名称/处罚时间/被处罚银行的经纬度。

## 28. 地方政府留言板文本数据

- 地方政府留言板问答文本内容全面抓取，时间跨度近10年（2011.01.01-2019.12.31），数据量级过百万，且未来仍会持续更新；
- 字段：问题及回复内容/点赞数/版块/用户昵称等。

## 29. CCTV新闻联播文本数据

- 文本类非结构化数据，字段简洁、内容丰富，时间区间为2016.06.15至2019.12.31，并持续更新；
- 字段：发布日期/ID列/新闻标题及内容。

## 30. 谣言数据-open

- 较真查证平台上的辟谣信息，时间区间为2006.04至2020.02，内容丰富；
- 字段：谣言ID/标题/内容/结论/解释源/作者/发布时间等。

## 31. 唐诗宋词数据-open

- 包含诗词表和作者信息表等六张表格，最多达25万行；
- 字段：唐诗、宋词、宋诗作者/名/正文等。





# CnOpenData 数据目录

## 32. 空气质量站点监测数据-open

- 各城市空气质量的日度数据、位置信息、监测数据（自2014年5月13日起）；
- 字段：监测点编码/名称/城市/经纬度/数据采集日期/时刻/城市/空气质量指数。

## 33. 中国历史天气数据-open

- 中国所有城市（含县）自2011年以来的历史天气数据，指标包含温度、天气、风向风力、空气质量（自2016起）等内容；
- 字段：省份/区县/日期/最低最高温度/风向/空气质量指数/空气质量登记。

## 34. 台风数据

- 含台风基本信息、台风路径信息、台风风圈半径表；
- 字段：台风编号/中英名称/发生日期/台风强度/经纬度/风俗/移动方向/路径编号/风速级别等。

## 35. 机场信息数据-open

- 含世界机场数据、中国机场数据（2020.01.15之前）；
- 字段：机场名/经纬度/机场地址/电话/开通年度月度；列车编号/起始站编号名称/出发时间/车站编码/服务类型/途径站名称/高铁站开工及通车时间/高铁线路。

## 36. 铁路信息数据

- 含12306中所有列车时刻信息（2007、2011、2013年后每年）、所有高铁线路表及各高铁站点开通时间表。
- 字段：列车编号/起始站编号名称/出发时间/车站编码/服务类型/途径站名称/高铁站开工及通车时间/高铁线路。

## 37. 中国电影票房数据

- 含电影基本信息表及票房详细信息表（2019.10.15前）；
- 字段：电影名/上映年度/类型/累计票房/导演及主演/单周及累计票房等

## 38. 电影评分数据

- 全面涵盖了电影相关的所有信息，尤其包括大量简洁的短篇以及优质的长篇电影评论。
- 字段：电影名称/豆瓣评分/评分ID/评分/评论内容/演员名称/出生日期及出生地等；

## 39. 国际足球比赛结果数据-open

- 1872年至2017年间的足球赛事结果数据。
- 字段：比赛日期/主队及客队名字/比赛名称/全队得分（不包括点球）。

## 40. 失信被执行人数据

## 41. 股吧评论数据

## 42. 中国各省份党媒报纸文本数据

数据不断更新中……

气象数据

交通数据

娱乐数据



# 数据的顶端 也是学术的顶端

## 服务内容

### 线上平台已发布的数据

### 与客服联系获取私有数据

部分私有数据已经在国内外顶级期刊上发表多篇学术论文。

### 个性化的数据定制服务

这是我们的老本行，您只需说明需求，我们评估后便可进行抓取。  
数据量越大、抓取难度越高的数据，我们越有信心！



Q CnOpenData

公众号



Q CNopendata-ybl

销售经理 - 杨宝璐



Q CnOpenDataOrg

机构服务

官网链接：[www.cnopendata.com](http://www.cnopendata.com)

邮箱地址：[yangbaolu@cnopendata.com](mailto:yangbaolu@cnopendata.com)