## 我国开放数据元数据领域的“词汇表”调查

### 1. 数据描述

本次分析的数据源自中国部分省级行政区的开放数据平台。数据包括不同省级行政区的分类词汇表。使用DBSCAN算法对这些数据进行了聚类分析，并生成了热图和词云图来展示各区域词汇表的异同。

### 2. 聚类结果

尝试使用K-Means聚类，层次聚类和DBSCAN聚类算法，根据数据集主题词汇表，对各省级行政区进行聚类。经人为观察绘制出的热图判定，DBSCAN聚类算法的结果最具有可解释性，简单调参后得到最终结果：将省级行政区分为三个主要类别和一些噪音点。每个类别中的省级行政区在词汇表上具有一定的相似性。

### 3. 聚类结果可视化

分别为每个聚类类别绘制了词汇表热图，并绘制了中国地图，展示各省级行政区的聚类结果。以下是各个聚类的分析：

##### 聚类 0

* **主题**：财税金融, 信用服务, 生活服务, 经济建设, 农业农村, 社会保障, 劳动就业, 生活安全, 餐饮美食, 环境与资源保护, 消费购物, 医疗健康, 企业服务, 交通服务, 宗教信仰, 旅游住宿, 文体娱乐, 教育科研, 政府机构与社会团体, 房屋住宅
* **特征**：该聚类中的省级行政区在特定的词汇上具有较高的相似性。主要集中在一些常用分类，如经济建设、教育科技、交通服务等。
* **省级行政区**：上海市、广东省、江西省。

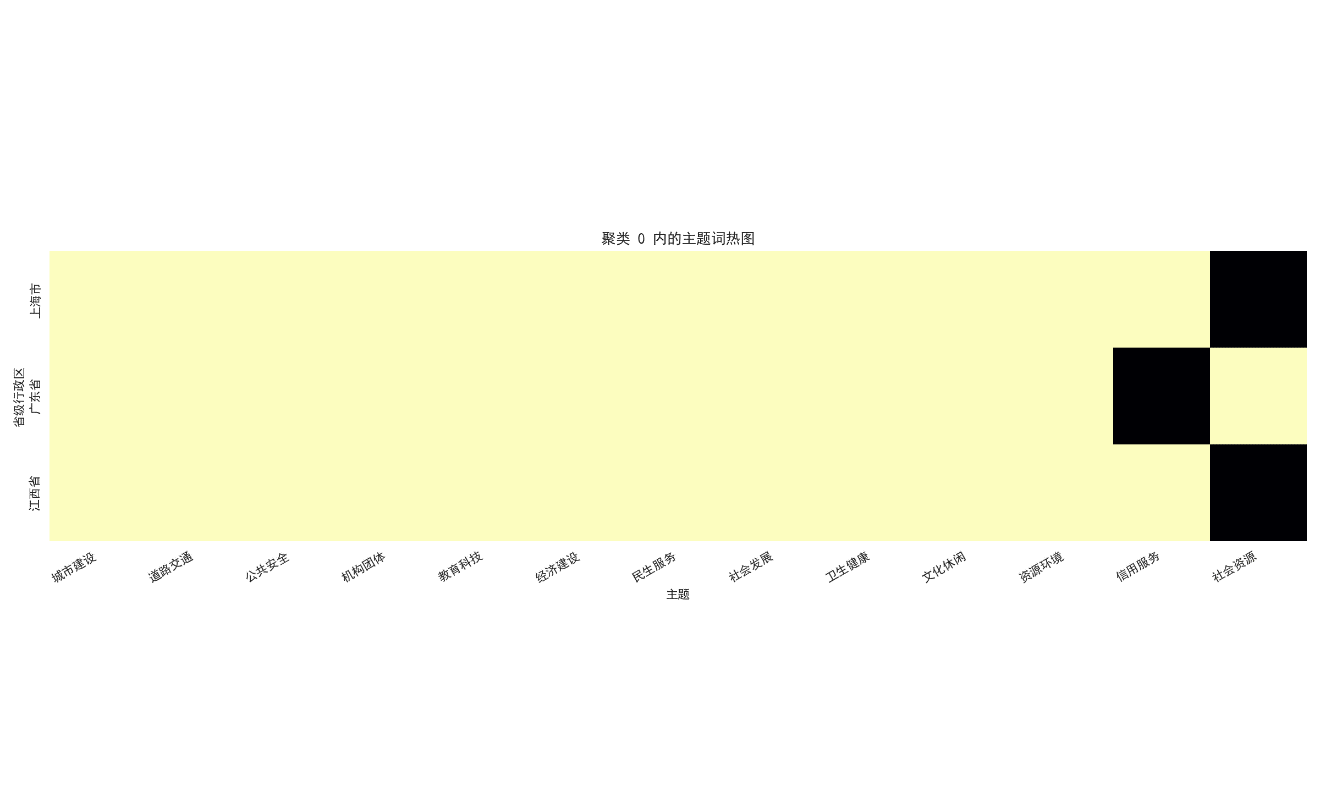


Figure 1 聚类0内的主题词热图

##### 聚类 1

* **主题**：机构团体, 网络安全, 公共安全, 社保就业, 地理空间, 城建住房, 安全生产, 教育文化, 气象服务, 法律服务, 农业农业, 交通运输, 信用服务, 环保主题, 其他, 气象主题, 市场监管, 科技创新, 财税金融, 社会救助, 生态环境
* **特征**：该聚类包含了多个省级行政区，并在某些特定类别上有较多的重叠。比如在社会保障、文化旅游、资源环境等词汇上具有相似性。这些省份在各自领域的数据开放较为全面，可能是由于地域和产业多样性较大，导致主题词汇分布广泛。
* **省级行政区**：内蒙古自治区、吉林省、四川省、天津市、安徽省、广西壮族自治区、浙江省、海南省、贵州省、重庆市。

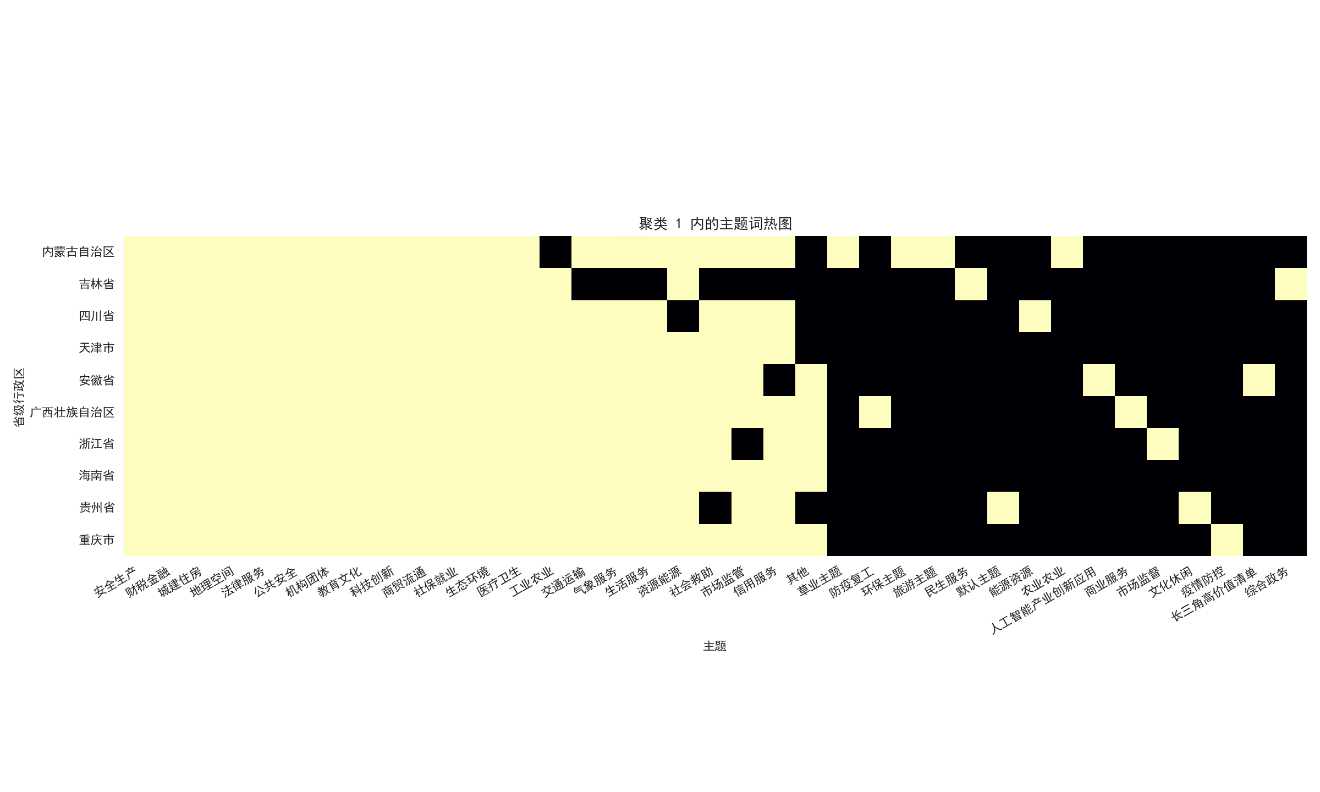


Figure 2聚类1内的主题词热图

##### 聚类 2

* **主题**：机构团体, 卫生健康, 社会发展, 经济建设, 综合政务, 农业农村, 市场监管, 安全生产, 资源环境, 文化休闲, 新冠疫情, 公共安全, 气象服务, 城建住房, 社保就业, 信用服务, 财税金融, 道路交通, 社会救助, 生态环境
* **特征**：该聚类中的省级行政区在工业农业、社会发展、科技创新等类别上有较高的相似性。
* **省级行政区**：山西省、山东省。

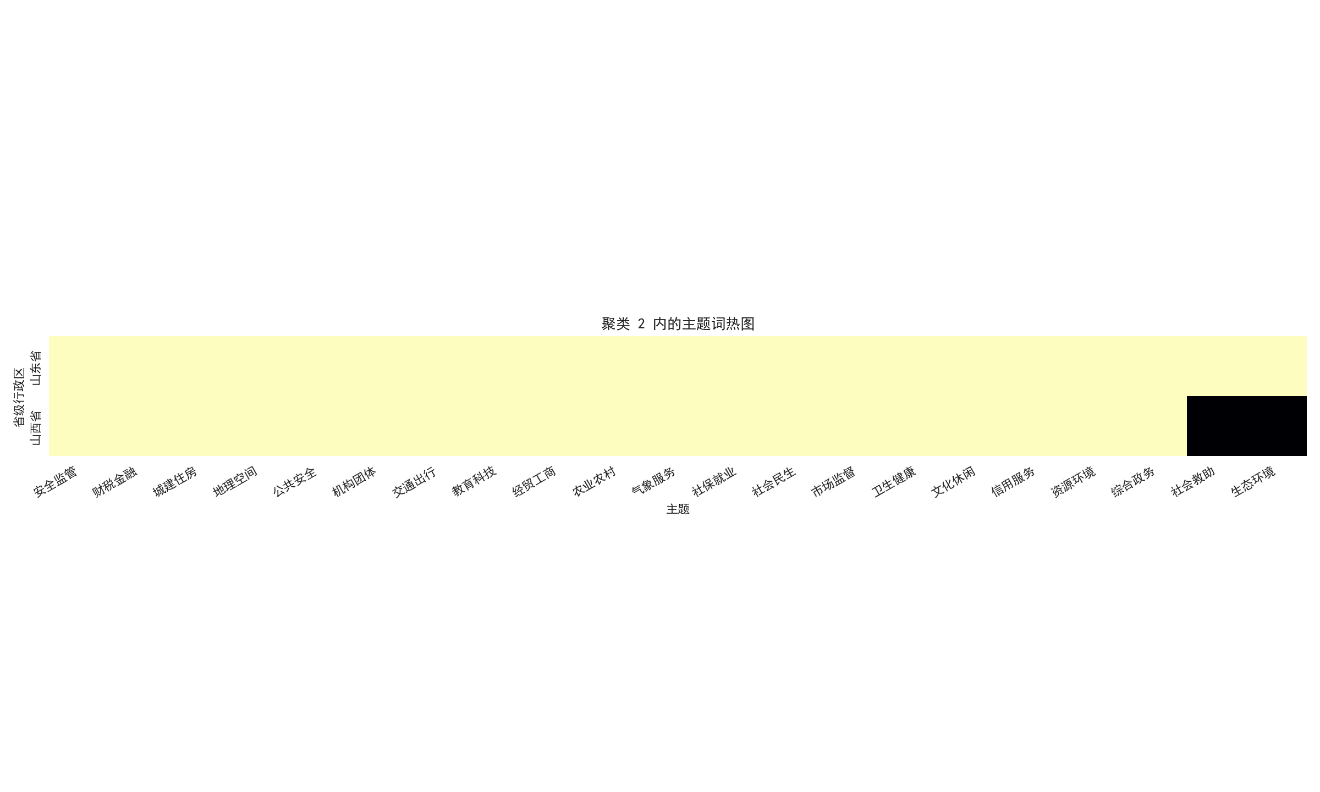


Figure 3聚类2内的主题词热图

##### 中国省份聚类结果地图

* **地图说明**：地图展示了每个省级行政区的聚类结果，不同颜色代表不同的聚类类别，如绿色区域代表噪音点。

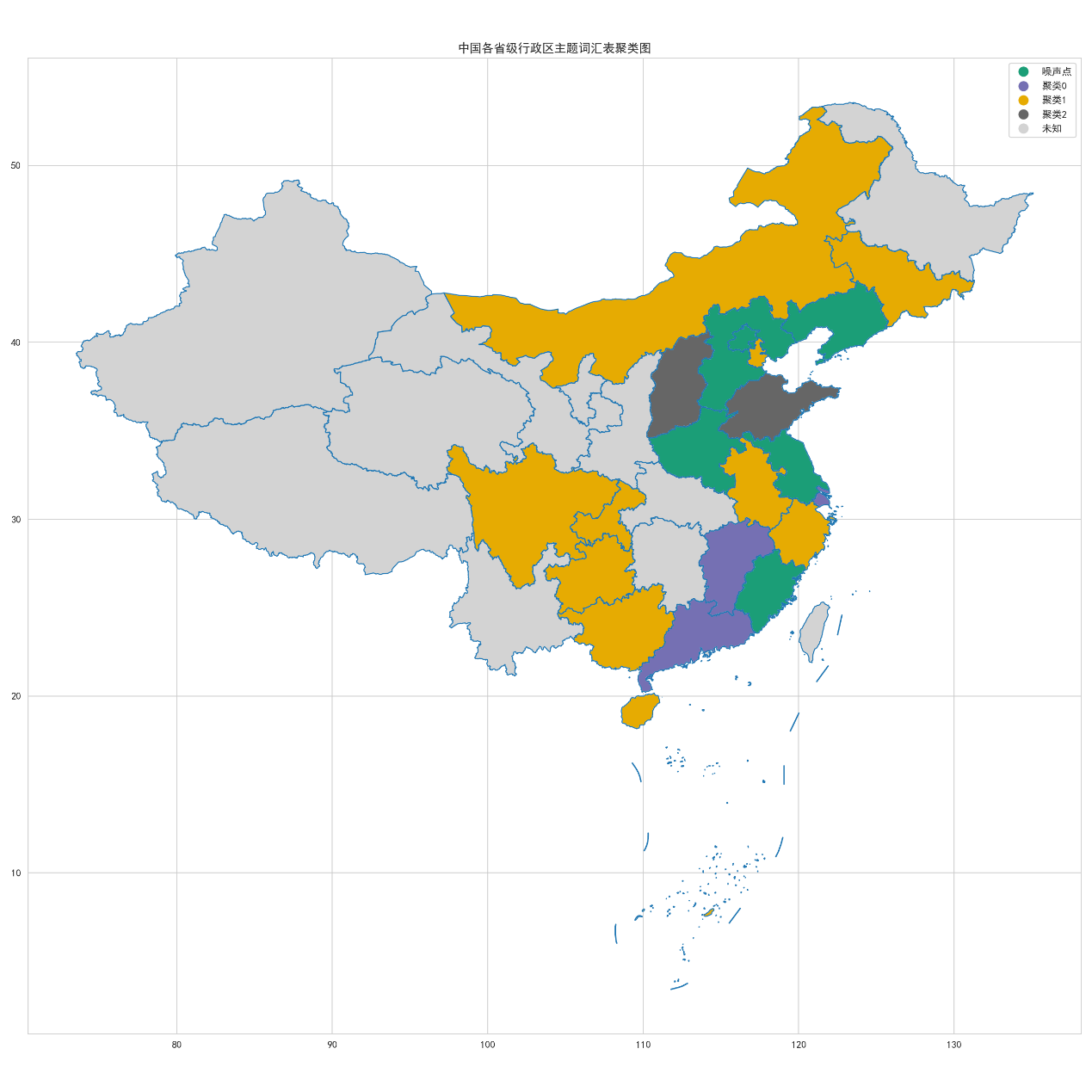


Figure 4中国各省级行政区主题词汇表聚类图

### 4. 噪音点详细分析

针对DBSCAN聚类算法检测到的噪音点，具体分析这些省级行政区的词汇表特点，尝试找出它们在分类上的独特之处。

#### 北京市

**主题词汇表**： 财税金融, 餐饮美食, 房屋住宅, 环境与资源保护, 交通服务, 教育科研, 经济建设, 劳动就业, 旅游住宿, 农业农村, 企业服务, 社会保障, 生活安全, 生活服务, 文体娱乐, 消费购物, 信用服务, 医疗健康, 政府机构与社会团体, 宗教信仰

**分析**：

* 北京市的分类词汇表非常全面，涵盖了从基础设施、经济建设到社会保障、文化娱乐等多个方面。
* 特别是包括了宗教信仰、餐饮美食、企业服务等类别，这些类别在其他省级行政区中不常见，显示了北京作为首都在数据分类上的全面性和多样性。

#### 江苏省

**主题词汇表**： 安全生产, 出境入境, 工程建设, 行业准营, 环境资源, 交通旅游, 教育培训, 劳动就业, 纳税缴费, 其他, 三农服务, 设立变更, 社会保障, 涉外服务, 生活服务, 投资立项, 文物保护, 医疗卫生, 执业资格, 住房保障

**分析**：

* 江苏省的词汇表也较为全面，特别是出境入境、涉外服务、纳税缴费、执业资格等类别显示了其在外贸、涉外服务和专业资格认证上的数据开放。
* 设立变更、三农服务、文物保护等分类表明其在农业服务和文化保护方面的独特数据分类。

#### 河北省

**主题词汇表**： 安全监管, 财税金融, 城建住房, 道路交通, 地理空间, 机构团体, 教育科技, 社保就业, 食药安全, 卫生健康, 无, 信用服务, 资源环境

**分析**：

* 河北省的词汇表比较均衡，涵盖了从基础设施建设、教育科技到社会保障和健康服务等方面。
* 包含了“无”这一分类，可能是特殊的数据标记或缺失数据的分类。

#### 河南省

**主题词汇表**： 财税金融, 城建住房, 工业农业, 公共安全, 公共数据, 机构团体, 交通运输, 教育文化, 经济发展, 科技创新, 民生服务, 其他, 气象服务, 社保就业, 社会发展, 审计监督, 市场监督, 通信, 文化休闲, 医疗卫生, 资源环境, 自然资源

**分析**：

* 河南省的分类非常全面，特别是公共数据、气象服务、审计监督和市场监督等类别显示了其在数据开放和监管上的特殊分类。
* 自然资源、经济发展、科技创新等类别表明其在资源管理和科技发展的侧重点。

#### 福建省

**主题词汇表**： 安全生产, 财税金融, 城乡建设, 地理空间, 法律服务, 工业农业, 公共安全, 公共服务, 海洋服务, 机构团体, 交通运输, 教育科技, 闽台和海外合作, 气象服务, 商贸流通, 社保就业, 社会救助, 生活服务, 生态环境, 市场监管, 统计服务, 卫生健康, 文化旅游, 信用服务, 医疗保障, 资源能源

**分析**：

* 福建省的分类词汇表独特地包括了海洋服务、闽台和海外合作，显示了其在海洋资源和对台合作方面的数据分类。
* 包含了法律服务、公共服务、社会救助等类别，表明其在社会服务和法律保障上的分类细化。

#### 辽宁省

**主题词汇表**： 安全监管, 地理, 交通运输, 教育, 就业, 科技, 农业, 企业登记监管, 气象, 社会保障, 生态, 统计, 卫生, 文化, 信用, 医疗, 资源

**分析**：

* 辽宁省的词汇表覆盖了基础设施、科技、农业、企业监管等多个方面。
* 包含企业登记监管、气象、生态和统计等类别，表明其在企业管理、环境监测和数据统计上的分类特色。

### 5 词云图展示

这张词云图展示了聚类前所有省级行政区的词汇表。词云图通过不同大小和颜色的文字来表示词频的高低，词频越高的关键词，字体越大。



Figure 5聚类前所有省级行政区主题词云图

### 6. 结论

* **聚类 0**：词汇表多为基础设施建设和经济相关类别，可能表明这些省级行政区更注重基础设施和经济建设。
* **聚类 1**：词汇表中涉及到的类别较为多样，可能是由于地域和产业多样性导致的。
* **聚类 2**：词汇表更多集中在工业、农业和科技创新方面，可能反映了这些省级行政区在这些领域的政策和数据开放。
* **噪音点**：这些噪音点显示了各省级行政区在数据分类上的独特性，可能由于其独特的经济、地理和社会背景，使得它们未能归类到任何一个主要聚类中：
  + **北京市**：分类词汇表非常全面，涵盖了从基础设施、经济建设到社会保障、文化娱乐等多个方面，显示了其作为首都在数据分类上的全面性和多样性。
  + **江苏省**：分类上注重外贸、涉外服务和专业资格认证，农业服务和文化保护也有独特的分类。
  + **河北省**：分类均衡，特别是“无”这一分类可能是特殊的数据标记。
  + **河南省**：分类非常全面，特别在公共数据、气象服务、审计监督和市场监督等方面有特殊分类。
  + **福建省**：分类独特，特别是海洋服务和闽台合作显示了其区域特色。
  + **辽宁省**：分类覆盖基础设施、科技、农业和企业管理，显示了其在环境监测和数据统计上的特色，同时也是唯一一个多使用二字主题的省级行政区。
* **词云图：**从词云图中可以看出，**公共安全**、**机构团体**、**社保就业**、**财税金融**、**地理空间**、**生态环境、信用服务**是最核心的分类。这些分类在各省级行政区的数据开放中具有普遍重要性。一些较为特殊的分类，如**闽台和海外合作**、**宗教信仰**和**企业服务**等，显示了某些地区的独特需求和关注点。这些分类在词云图中字体较小，但也反映了数据开放的多样性和地方特色。

### 附录

表 我国除港澳台外，各省级行政区公共数据平台建设汇总表（最后测试时间2024-05-07）

| 序号 | 省份 | 平台名称 | 平台链接 | 备注 | 最近更新时间 |
| --- | --- | --- | --- | --- | --- |
| 1 | 北京市 | 北京市公共数据开放平台 | <https://data.beijing.gov.cn/> | 无开放类型，下载均需登录，按主题、机构 | 2024-05-07 |
| 2 | 天津市 | 天津市信息资源统一开放平台 | <https://data.tj.gov.cn/> | 无条件开放下载需登录，按主题、市级、区级、综合得分 | 2024-04-30 |
| 3 | 河北省 | 河北省公共数据开放网 | <http://hebdata.hebyun.gov.cn/> | （资源目录）大概率无法下载（提示注册），按主题、按部门 | 2023-11-29 |
| 4 | 山西省 | 山西省公共数据开放网 | <http://data.shanxi.gov.cn/> | **无法下载**，按省直部门、领域筛选、开放类型 | 2023-12-29 |
| 5 | 内蒙古自治区 | 内蒙古自治区公共信息资源开放平台 | <http://open.nmgdata.org.cn/> | 下载需登录，按数据提供单位、主题、开放方式 | 2024-04-12 |
| 6 | 辽宁省 | 辽宁省公共数据开放平台 | <https://data.ln.gov.cn/> | 无条件开放直接下载，有条件开放申请，按领域、格式、开放类型、省直部门、地市、企事业单位，无障碍浏览 | 2024-03-22 |
| 7 | 吉林省 | 吉林·数据开放 | <https://data.jl.gov.cn/> | 登陆下载，按数据提供部门、主题分类、数据格式 | 2024-04-25 |
| 8 | 黑龙江省 | null | null | null | null |
| 9 | 上海市 | 上海市公共数据开放平台 | <https://data.sh.gov.cn/> | 有条件开放申请，无条件开放直接下载，按数据领域、资源类型数据、提供单位、综合得分、其他数据 | 2024-05-07 |
| 10 | 江苏省 | 江苏省公共数据开放平台 | <http://data.jszwfw.gov.cn:8118/extranet/> | 下载需登录，按省级部门、设区市、所属主题、所属行业、资源类型、开放类型；无障碍浏览，老年人专区 | 2024-05-07 |
| 11 | 浙江省 | 浙江·数据开放 | <https://data.zjzwfw.gov.cn/> | 下载需登录，按数据来源单位、领域、标签、格式；网站无障碍，进入老年模式 | 2024-05-07 |
| 12 | 安徽省 | 安徽省数据开放平台 | <http://data.ahzwfw.gov.cn:8000/dataopen-web/index.html> | 下载需登陆，按开放属性、所属领域、所属区域、提供部门 | 2024-01-09 |
| 13 | 福建省 | 福建省公共数据资源统一开放平台 | <https://data.fujian.gov.cn/>‍ | 下载需登录，按领域、格式、开放类型；无障碍浏览，长者模式 | 2024-01-31 |
| 14 | 江西省 | 江西省公共数据开放平台 | <https://data.jiangxi.gov.cn/> | 下载需登录，按部门、领域、开放类型、责任清单 | 2024-04-28 |
| 15 | 山东省 | 山东公共数据开放网 | <https://data.sd.gov.cn/> | 无条件开放直接下载，有条件开放申请，按领域、格式、开放类型、省直部门、地市、企事业单位、综合得分；无障碍通道 | 2024-05-07 |
| 16 | 河南省 | 河南省公共数据运营服务平台 | <http://hndataops.com/> | 只有几十个数据接口，按数据来源、行业领域、服务类型 | 2024-03-08 |
| 17 | 湖北省 | null | null | null | null |
| 18 | 湖南省 | 湖南政务大数据公众门户 | <https://data.hunan.gov.cn/etongframework-web/> | 建设中 | 2024-03-28 |
| 19 | 广东省 | “开放广东”全省政府数据统一开放平台 | <https://gddata.gd.gov.cn/> | 下载需登录，按部门、主题分类、格式、开放模式，无障碍浏览 | 2024-05-06 |
| 20 | 广西壮族自治区 | 广西壮族自治区公共数据开放平台 | <https://data.gxzf.gov.cn/> | 下载需登录，按区直部门、地市、企事业单位、领域、格式、数据格式、开放类型；长者专区 | 2024-04-23 |
| 21 | 海南省 | 海南省政府数据统一开放平台 | <https://data.hainan.gov.cn/> | 下载需登录，按主题、部门、综合得分；网站无障碍，适老版 | 2024-05-07 |
| 22 | 重庆市 | 重庆市公共数据开放系统 | <http://data.cq.gov.cn/> | 下载需登录申请，按数据部门、区县、行业、主题、格式、开放属性 | 2024-05-07 |
| 23 | 四川省 | 四川公共数据开放网 | <https://www.scdata.net.cn/> | 无条件开放直接下载，有条件开放申请，按领域、格式、开放类型、省直部门、地市 | 2024-04-11 |
| 24 | 贵州省 | 贵州省政府数据开放平台 | <http://data.guizhou.gov.cn/> | 下载需登录并订阅，按主题、场景、部门、类型、格式 | 2024-05-06 |
| 25 | 云南省 | 云南省公共数据平台 | <http://106.227.94.96:8894/> | 正在建设中 | null |
| 26 | 西藏自治区 | null | null | null | null |
| 27 | 陕西省 | null | null | null | null |
| 28 | 甘肃省 | null | null | null | null |
| 29 | 青海省 | null | null | null | null |
| 30 | 宁夏回族自治区 | 宁夏公共数据开放平台 | <https://opendata.nx.gov.cn/portal/index> | 名存实亡，没有一个能下载的 | 2022-06-23 |
| 31 | 新疆维吾尔自治区 | null | null | null | null |