# 采集途径、采集行为、使用目的

## 数据的采集途径

通过什么途径爬取数据，这个是最需要重视的一点。总体来说，未公开、未经许可、且带有敏感信息的数据，不管是通过什么渠道获得，都是一种不合法的行为。

所以在采集这类比较敏感的数据时，最好先查询下相关法律法规，特别是用户个人信息、其他商业平台的信息 等这类信息，寻找一条合适的途径。

### 个人数据

采集和分析个人信息数据，应该是当下所有互联网都会做的一件事，但是大部分个人数据都是非公开的，想获得必须通过合法途径，可参见『网络安全法』第四十一条：

网络运营者收集、使用个人信息，应当遵循合法、正当、必要的原则，公开收集、使用规则，明示收集、使用信息的目的、方式和范围，并经被收集者同意...

也就是必须在提前告知收集的方式、范围、目的，并经过用户授权或同意后，才能采集使用，也就是我们常见的各种网站与 App 的用户协议中关于信息收集的部分。

### 公开数据

从合法公开渠道，并且不明显违背个人信息主体意愿，都没有什么问题。但如果通过破解、侵入等“黑客”手段来获取数据，那也有相关法律等着你：

刑法第二百八十五条第三款规定的“专门用于侵入、非法控制计算机信息系统的程序、工具”：

（一）具有避开或者突破计算机信息系统安全保护措施，未经授权或者超越授权获取计算机信息系统数据的功能的；

违反 Robots 协议

虽然 Robots 协议没有法规强制遵守，但 Robots 协议作为行业约定，在遵循之下会给你带来合法支持。

因为 Robots 协议具有指导意义，如果注明 Disallow 就说明是平台明显要保护的页面数据，想爬取之前应该仔细考虑一下。

## 数据的采集行为

使用技术手段应该懂得克制，一些容易对服务器和业务造成干扰甚至破坏的行为，应当充分衡量其承受能力，毕竟不是每家都是 BAT 级。

### 高并发压力

做技术经常专注于优化，爬虫开发也是如此，想尽各种办法增加并发数、请求效率，但高并发带来的近乎 DDOS 的请求，如果对对方服务器造成压力，影响了对方正常业务，那就应该警惕了。

如果一旦导致严重后果，后果参见：

《刑法》第二百八十六条还规定，违反国家规定，对计算机信息系统功能进行删除、修改、增加、干扰，造成计算机信息系统不能正常运行，后果严重的，构成犯罪

所以请爬取的时候，即使没有反爬限制，也不要肆无忌惮地开启高并发，掂量一下对方服务器的实力。

### 影响正常业务

除了高并发请求，还有一些影响业务的情况，常见的比如抢单，会影响正常用户的体验。

## 数据的使用目的

数据使用目的同样是一大关键，就算你通过合法途径采集的数据，如果对数据没有正确的使用，同样会存在不合法的行为。

### 超出约定的使用

一种情况是公开收集的数据，但没有遵循之前告知的使用目的，比如用户协议上说只是分析用户行为，帮助提高产品体验，结果变成了出售用户画像数据。

还有一种情况，是有知识产权、著作权的作品，可能会允许你下载或引用，但明显标注了使用范围，比如不能转载、不能用于商业行为等，更不能去盗用，这些都是有法律明文保护，所以要注意使用。

其他情况就不列举了。

### 出售个人信息

关于出售个人信息，千万不要做，是法律特别指出禁止的，参见：

根据《最高人民法院 最高人民检察院关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》第五条规定，对“情节严重”的解释：

（1）非法获取、出售或者提供行踪轨迹信息、通信内容、征信信息、财产信息五十条以上的；

（2）非法获取、出售或者提供住宿信息、通信记录、健康生理信息、交易信息等其他可能影响人身、财产安全的公民个人信息五百条以上的；

（3）非法获取、出售或者提供第三项、第四项规定以外的公民个人信息五千条以上的便构成“侵犯公民个人信息罪”所要求的“情节严重”。

此外，未经被收集者同意，即使是将合法收集的公民个人信息向他人提供的，也属于刑法第二百五十三条之一规定的“提供公民个人信息”，可能构成犯罪。

### 不正当商业行为

如果将竞品公司的数据，作为自己公司的商业目的，这就可能存在构成不正当商业竞争，或者是违反知识产权保护。