###### 裁判文书数据清理发掘规划

2021年6月29日稿，by 王鑫。

可以开展如下工作：

1. 将数据按照“案件类型+省份+裁判年份”分为若干个数据集。
2. 对一些标准的变量进行清理，包括：案件类型、案件类型代码、裁判日期、法院名称、法院代码、地区、地区代码、审判程序。
3. “案号”中是否能提取出额外信息？是否可以作为案件的唯一指代？
4. 从标题中提取“案由”，基于案由可以对案件类型进行进一步分类。
5. 选定一类案件（比如“知识产权纠纷”），对正文内容进行清理，包括如下工作：
   1. 提取原告信息。
   2. 提取被告信息。
   3. 提取原告诉求。
   4. 提取被告诉求。
   5. 提取判决内容。
   6. 提取审判员等相关信息。是否可以匹配外部数据，比如法官的个人信息？
   7. 提取如案件发生时间等其他信息。
   8. 如二审及之后程序，识别关联一审案号。