1. 总体流程
2. 更新数据列表，手动筛选数据来源，
3. 筛选和分级数据
4. 下载筛选过的新增加的数据
5. 根据需求分发数据（查重）并记录
6. 对于各类有问题的、回收的数据记录与处理
7. 具体介绍
8. 更新：

目前每隔一个月更新一次数据列表，更新参数有：更新起始日期、上一次产生的列表、杂志期刊目录、本次更新时间。

一般不扩增杂志种类的情况下，仅需更改脚本中起始日期、上一版数据列表名

更新后会在上一版列表后加上这次的数据，这也将会成为下次更新的旧列表输入。和上一版列表对比，把新加的后面一截手动复制到工作表上，不要在List上直接修改，List只在更新中当模板，更新的内容复制到工作表上进行后续处理记录，结束更新。

1. 筛选：

对于每次新加的数据，有的有GSE号，有的没有。需要做的是：

对于新增的每一篇每一篇文章，根据文章号，打开文章，阅览查找其数据存在哪里、编号什么，我们只要存在GEO和EBI上的非原始数据（已经有GSE号的也要查找，核对，有的会出错），并把数据集编号记在表中（子系列只需要记它们的爸爸就行，下载器会把爸爸下面所有的儿子都下下来）。

同时在每一个数据后记下它的难度，很容易，只需要在查找时看看有多大，一般分为几十兆，四百兆以下，一个G以下，和更大。下载优先下小的，分发策略待会说。

记录下GSE、EBI号和难度后，其他不符合要求和只有原始数据的文章就不用管了，将有GSE号的数据用于下载，EBI的数据不用下载，记录好待发就行。

1. 下载：

下载也是自动化脚本，对于需要下载的GSE，在下载前，在dataset\_list和分发记录中查重一下，我们下载的一定是：单细胞转录组且不重复的！，确定没问题后，在脚本对应位置填上编号，运行脚本nohup python3 test\_downloading.py &，可用htop查看进程情况或终止进程。

下载完成后，需要查看每个数据集是否下载了，下载的里面有没有下载失败的，有数据的有没有漏了文件的，确定无误后，下载成功。

接下来，对数据按照难易程度分到四个文件夹中，并记录待发列表中。

1. 分发：

分发的数据包括下载好的待分发的GSE数据和已审查过的EBI数据：

来了一个人，要数据，根据他的水平，依次分发难易不等的GSE数据。每分发一个数据，需要在服务器脚本、每周分发记录、总分发表、工作表中同时将该数据状态改为已分发！不可缺漏！，分发就移到他目录下就行。

对于老手，做的不错的，才分EBI，只告诉他们编号，让他们自己下载，按照流程自己建文件夹然后处理。

然后每周结束将分发记录备份到GOOGLE中

1. 其他：

有时候数据要回收，就挪回对应难度文件夹，所有记录表状态改为待分发或回收待分发即可，之后正常分发就行；

发的数据不符合要求，则在表上记录属于哪种问题即可；

扩充杂志类型等，需要对扩增出的数据进行查重，因为虽然原来检索的没有这些杂志名称，但有是这个杂志的数据被另一篇杂志发的文章用了，所以更新的杂志产生的新的列表需要仔细去重，无论如何，宗旨是：1.必须单细胞转录组，2.不能和以前重复

有问题随时联系：李金虎，15645112932，QQ：1412167773，微信：ljh1412167773

未来的工作：1，扩充杂志，2，扩充数据类型，单细胞全组学