针对目前完成的数据集我们主要检查三个部分，我会把检查过的数据集错误标注在“关于现有数据集全面核查情况的记录”中。下面我会列出每个数据集需要人工检查的部分，以及每一个错误对应数据集需要修改的内容：

1， 数据集的 unstructureData 部分

这一部分对应数据集展示的首页，需要检查的部分主要是展示出来的几个关键 keys，包括：

- **题名 metadata['title']**
  题名是否对应的是文章标题，而不是数据集标题
- **GSE 号 metadata['accessionNumber'] / PMID metadata['pubmedID']**
  这两部分是否是和数据及内容相对应的，正确的编号
- **摘要 metadata['abstract']**
  是否完整，不能复制多或者复制少了
- **摘要图 metadata['figureURL']**
  对于明确表示有 graphic abstract 的文章，我们需要把这张图放在展示页面；如果没有 graphic abstract，那么放文章第一张图。之前的数据集大部分人没有放图，或者放的是 cluster 图，这部分需要修改的比较多。当摘要图模糊时，更换链接，在文章页面访问原图，使用原图链接；或者访问杂志网站，使用杂志提供的图片链接
- **物种 metadata['taxonomyID']**
  大部分为人 Homo sapiens/鼠 Mus musculus，其余会显示 others，检查与文中所用实验对象是否一致
- **组织 metadata['tissue']**
  是否对应文中实验取材来源，以及是否是词表中包含的关键字；除了文中明确取材对象是胚胎 embryo 时可填 notAvailable，其他时候都不应该空，都应填入对应字段
- **建库方法 metadata['libraryPreparationMethod']**
  是否对应文中和数据库中的处理方法（一般在文中 method 和数据集 sample 中的 protocol 位置），以及拼写是否对应词表中的正确格式。
- **杂志 metadata['journal']/出版日期 metadata['publicationDate']**
  作者 metadata['authors']/关键词 metadata['keywords']
  这四部分由内置函数获取，一般不会出错。

2， 数据集的 cellAnnotation 部分

这一部分对应数据集展示的 Dimensional Reduction 中的 clusterName 及对于细胞的其余注释，对应 cellAnnotation 中的”meta_”一类字段。需要检查 clusterName 是否有误，分类项是否有意义且是否有缺漏

重点保留 tissue，sourceName，cellOntologyName，其余没有意义的字段（例如给每个细胞加上的编号）删去，内容格式混乱的需要调整。

3， 数据使用完整性

这一部分主要关注数据集有没有分 subDataset，有没有分对，以及对于提供的数据有没有全部使用上。

对于分了 subDataset 的数据集，除了每个部分都需要检查 unstructureData 展示部分是否正确以外，还需要额外检查 description，即对于分 part 标准的叙述。有些

subDataset 是基于不同物种中的实验，有些是基于不同的聚类实验，还有些是不同的处理。这些都应该在文中和数据库中找到对应的证据，否则应该算是分 part 有误。对于分 part 的错误，包括：1）应该分 part 但是没有分；2）不应该分 part 但是分了。此外，在检查中有发现部分数据集中混杂了非单细胞的数据，体现为细胞数量多于文中提到的分析用量，这种情况下应该删去非单细胞数据并重新运算。

提供的数据有没有全部使用上也分为两种情况：1）对于原始数据没有全部使用上，表现为细胞数缺漏，或者缺少 subDataset；2）对于提供的 normalize 矩阵没有使用上，表现为没有生成 normalize 矩阵（normalize 矩阵大小 51B），TPM 直接由 raw_counts 生成（查代码）。这两种情况较为严重，均需要重新制作。

4， not scRNA-seq 数据集

记录并提交报告，等待清除。