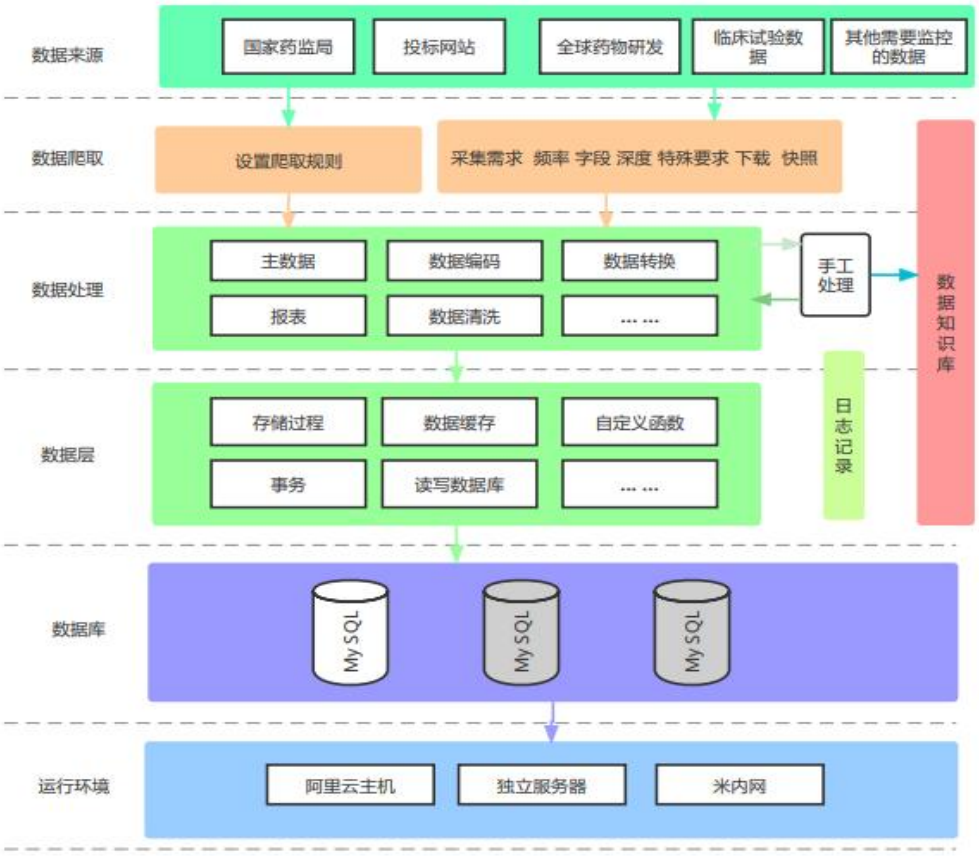
MN 工具型数据库需求调研报告

2022 年 9 月 15 日应产品部 MN 工具型数据库数据治理需求，实现米内网工具库审评进度（MED），上市药品（MID），中国临床试验三大模块及时数据更新，与产品部同事展开数据采集需求调研。



目前数据来源主要三个途径，第一，人工录入整理；第二通过“八爪鱼”等爬虫工具，由产品部编辑人员只做爬虫任务下载；第三，通过第三方公司提供爬虫服务。工作效率慢，无法及时抓取更新，部门数据来源获取困难。

调研会议上，CEO 张总就米内网现有的系统不足，以及未来对数字资产平台做出展望和规划如下：

1. 因各个数据系统之间，各自处理不同的数据，产生数据孤岛严重，无法实现各系统之间的数据互联互通，需要建立一套自己的数字资产管理平台，为米内网的数据更新，米内网的数据资产管理，保驾护航。

2. 目前网站，医院，药店各自都有一套属于自己的编码规则，各个规则没有相互关系，需要建立一套属于米内网的编码体系,服务与网站，医院，药店等终端。

3. 未来 6-8 个月数字科技能够研发一套及数据更新，数据爬取，数据清洗，数据处理，数据入库，数据发布的平台服务于产品部的 MN 工具型数据库数据更新平台。

4. 未来可以研发出一套系统类似中康 SIC 系统赋能平台，服务于医院，零售，药店的管理系统。

本次数据采集需求调研分成 5 个模块：国家药监局数据，临床试验数据，招投标数据，全球药物研发数据，其他需要监控的数据。

A. 国家药监局数据数据采集：

1. 国家药监局数据采集频率，一天/2 次。

2. 首次爬取全量的数据，后期需要检测更新，每隔一段时间，需要对数据做增量对比。

3. 审评任务公示中;灯泡需要转换成文字。

4. 受理品种信息需要根据评审任务公示进行更新(更新规则产品部细化）。

5. 优先审评公示 ，突破性治疗公示 双击详情获取。

6. 上市药品信息双击详情，相关详情 PDF 需要下载。

7. 药品目录跳转列表，详情页面的附件需要下载。

8. 药品送达信息存在多条，需要一一保存，形成评审信息流程。

9. 数据查询/药品栏目/医疗器械 这两个栏目数据需要产品部细化采集筛选栏目，整理提供相应的类库信息，以及现有爬取的成功的数据文件。

10. 一致性评价任务公示：新报,补充栏目数据都需要采集，灯泡图片需要转化成文字。

B. CDE 临床试验，WHO 临床试验数据采集：

1. CDE 临床试验，WHO 临床试采集频率，一天/1 次。

2. 采集栏目，搜索栏目全量数据。

3，列表目录双击，详情页面全部内容需要采集，文件需要下载。

C. 各省药品招投标网站平台数据采集

因各个招投标网站平台数据呈现类型不一致，而且药品只是招标品种的一个类目，而且需要有企业账号登，前期做以下内容数据的采集。

1. 各省药品招投标网站采集频率，一星期/1 次。

2. 采集内容：公告内容，公告链接，附件，页面镜像，标题，网站和二级链接镜像。

D. 全球药物研发信息数据

1. 全球药物研发信息网站采集频率，一天/1 次。

2. 采集栏目，数据字段要求，因国外网站原因，需要产品部整理具体采集要求

E. 其他需要监控的数据

1. 批签发库 采集频率，一月/2 次。

2. 批签发库首次爬取全量的数据，后期需要检测更新。

3. 上市公司业绩库 采集频率，半年/1 次。

4. 上市公司业绩库八爪鱼爬虫已经监控，已获取相关的数据

5. 医保局数据 采集频率，一月/1 次。

6. 医保局数据 需要获取省市两级联动的数据 。

7. 列表数据箭头点击查看详细信息。

为能尽快进行数据采集需求调研，下一阶段需求分析和需求评审。

下周二 会议红色字体类目，产品部需要补充完善。

各网站数据采集详细信息见数据采集网站需求V1.0.xls