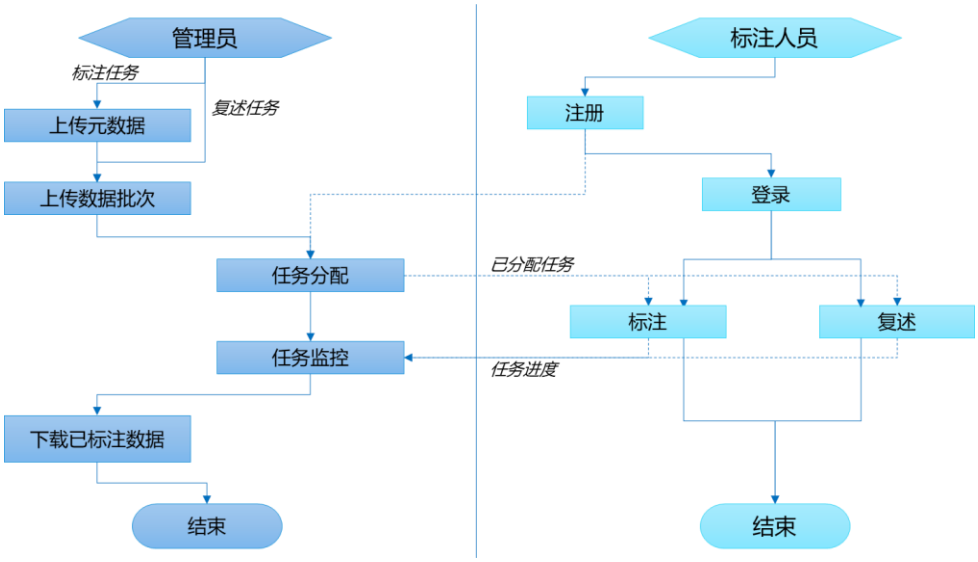
metaCAT 管理员操作指南

metaCAT 是一个专门为任务型对话数据而设计的开源 WEB 标注工具。它提供了基于元数据的全方位标注，覆盖了领域、意图、位置信息等对话要素的标注。同时，通过基于规则约束的实时标注检测机制来提升标注的质量。此外，metaCAT 还集成了 ASR 用于对话文本的录入，在提升对话复述采集的效率的同时，也引入了更丰富语言多样性和鲁棒性。

1 标注流程介绍

本标注工具的主要流程如下图所示：



其中管理员相关的流程要素有：

标注任务：基于已人工采集的对话数据，针对对话的意图和槽位进行人工标注。

复述任务：基于规则算法自动生成的标准对话，借助于键盘录入或 ASR 采集工具，同时对用户侧和系统侧的对话内容进行自然语言的口语化复述，同时对标准对话附带的部分槽位进行重新标注。

上传元数据：上传用于描述标注数据本体结构的数据文件，仅用于标注任务，具体文件格式详见工具附带的 sample 文件。

上传数据批次：上传待标注/复述的打包压缩数据文件，具体文件格式详见工具附带的 sample 文件。

任务分配：将系统解析过的数据文件按批次分配给已注册的标注人员。

任务监控：监控已分配任务的执行进度。

下载已标注数据：下载最新的部分完成或全部完成的标注或复述任务的结果数据。

2 各模块功能介绍

2.1 系统操作

系统操作模块主要提供初始化及数据的上传和下载功能。

初始化目录（①）：对系统侧的文件存储目录进行初始化，同时也对 MongoDB 数据库进行初始化。

导入元数据（②）：针对标注任务导入元数据，数据格式请参见工具附带的 sample 文件。

批次导入（③④）：针对标注任务和复述任务上传数据文件，目前支持两类数据文件格式：JSON 批次和 RAW 批次（仅用于标注任务），两者的数据格式请参见工具附带的 sample 文件。

批次导出（⑤⑥⑦）：针对进行中或已完成的标注任务或复述任务，导出结果数据文件，目前支持三类数据文件格式：对话标注批次、MultiWOZ 标注批次（已自动转换为 MultiWOZ 公开数据集的标注格式）、对话复述批次，三种导出格式均保持与导入格式的一致性，同时附带了相关的任务进度信息。

操作错误记录（⑧）：以上操作过程中如果出错，系统将返回相应的错误记录，便于管理员定位和解决。

2.2 数据管理

数据管理模块主要提供数据的展示以及任务的分配等功能，通常与系统操作模块协同。

数据集元数据（①）：展示与数据集对应的元数据。

查看元数据（②）：点击之后将展示 JSON 格式的元数据内容。

元数据简介（③）：介绍本工具所使用的元数据以及各元素之间相互依赖关系。

标注人员选择（④）：从已注册的标注人员中选择一个用于分配任务。

任务选择（⑤）：选择标注任务或复述任务用于分配给标注人员。

分配批次（⑥）：将已选中的批次任务分配给特定的标注人员。

2.3 标注进度

标注进度模块主要提供当前已分配标注任务和复述任务的进度展示功能。

2.4 系统设置

系统设置模块主要提供管理员密码修改功能以及各项功能选项的开关设置。

管理员密码修改（①）：提供系统管理员密码修改功能，当前管理员缺省密码请参见工具开源代码 GitHub 中相关说明。

开关-系统侧标注（②）：是否提供系统侧标注功能，缺省为打开。

开关-对话轮次删除（③）：是否允许删除对话轮次，缺省为关闭。

开关-对话文本新增（④）：是否允许新增对话轮次（附加到对话末尾），缺省为关闭。

开关-ASR 输入（⑤）：是否允许使用 ASR 输入功能，缺省为关闭。

3 修订记录

版本号	修订日期	修订纪要
1.0	2020/07/31	初稿。