ICS 35.020
CCS L 70

团　　体　　标　　准

T/BSIA 006—2022

信息技术　数据流程服务技术规范 第1部分：通用技术

Information technology—Specification for data process service—Part 1：General technology

2022-12-05 发布　　2022-12-06 实施

北京软件和信息服务业协会　发布
中国标准出版社　出版

目　　次

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则　第 1 部分:标准化文件的结构和起草规则》的规定起草。

本文件是《信息技术　数据流程服务技术规范》的第 1 部分。《信息技术　数据流程服务技术规范》已经发布了以下部分:

——第 1 部分:通用技术;

——第 2 部分:技术评价。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由北京软件和信息服务业协会提出并归口。

本文件起草单位:北京软件和信息服务业协会、北京中关村软件园发展有限责任公司、北京国际大数据交易有限公司、北京火山引擎科技有限公司、北京爱数智慧科技有限公司、北京金堤科技有限公司、北京云测信息技术有限公司、北京百分点科技集团股份有限公司、杭州曼孚科技有限公司、百融云创科技股份有限公司、联科云创(北京)科技有限公司、京北方信息技术股份有限公司、京东科技控股股份有限公司、昆仑海比(北京)信息科技有限公司、蚂蚁科技集团股份有限公司、北京深度搜索科技有限公司、辽宁京数云大数据科技有限公司、北京三快在线科技有限公司、五八同城信息技术有限公司、北京鼎兴达信息科技股份有限公司。

本文件主要起草人:龙飞、张磊、仓剑、杨楠、邓延嵘、汪蔚、郎佩佩、张锐、张凯悦、金亮、闵楠、沈苏、罗磊、王潇蔓、王猛猛、温士苇、仝仕京、贾宇航、冯晨旭、马伟凯、寇蕾蕾、张隽宁、张韶峰、李金伟、朱勇、宁平、王义刚、陈昊天、吴利、刘经梅、王昌钰、孙兆琳、昌文婷、杜霖、童玲、刘吉、胡成锴、宿博、赵冰洁、罗华丽,时圣师。

引　言

根据《国家标准化管理委员会、中央网信办、国家发展改革委、科技部、工业和信息化部关于印发〈国家新一代人工智能标准体系建设指南〉的通知》(国标委联〔2020〕35号)和《北京市数字经济全产业链开放发展行动方案》要求,为进一步发挥北京软件和信息服务业协会行业促进作用,引导软件和信息服务企业响应国家和产业发展需要,促进数字经济产业链开放发展,推动数据流通、应用,结合北京软件和信息服务业实际,特制定本文件。

本文件依据国家对人工智能标准体系,信息技术服务、数据分类、数据质量评价等相关标准文件,以及软件行业协会服务企业的成功经验,对数据流程服务的技术规范、评价体系、评估机构以及监督要求提出了规范性标准。为用户采购和选择数据流程服务供应商提供参考,也作为数据流通交易的重要参考标准,并且为服务商内部服务设计和质量控制提供参考。

本文件是由相关软件行业协会、企业、评价机构、认证机构基于市场和行业发展需要而共同制定,有利于发挥行业自律和示范作用,促进软件产业健康、可持续发展,实现对客户的满意度。

信息技术　数据流程服务技术规范
第1部分:通用技术

1　范围

本文件规定了数据流程服务技术的总则、服务场景及其技术准则,以及结构化数据和非结构化数据流程服务技术。

本文件适用于:

a)　数据流程服务需求方采购数据流程服务时,对数据和服务产品及其供应商进行评价;

b)　从事数据流程服务、销售数据产品的企业或机构,建立数据和服务产品技术规范;

c)　从事数据资产评估、数据交易服务的企业和机构,建立数据和服务产品交易规则、规范;

d)　政府相关管理部门、产业园区等对数据流程服务进行事中、事后监管核查;

e)　其他需要应用的场合。

数据流程服务需求方、数据流程服务方、数据交易所、行业协会及行业管理部门相关业务参照使用。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 36344—2018　信息技术　数据质量评价指标

3　术语和定义

下列术语和定义适用于本文件。

3.1

数据流程服务　data process service;DFS

使用数字技术,从实体世界或信息系统中采集、获取数据,并按后续应用、流通要求处理、输出数据,围绕数据处理流程形成的一系列服务。

3.2

数据采集和预处理　data acquisition and preprocessing

采集事实、概念或指令等对象信息,形成原始数据,并对其进行处理,以保证数据质量达到后续使用的规范性要求。

3.3

数据分析集成　data analysis integration

分析多源数据,进行集成融合,以满足后续应用和服务的规范性要求。

3.4

数据标注　data annotations

通过标记、注释等工作,对数据进行处理,提取对象的特征,以保证数据质量达到后续数字应用使用的规范性要求。

3.5

内容审核　content review

对互联网用户上传、发布或共享的内容(文字,图片,音频,视频等数据)进行识别,通过标记、注释等工作,提取违反相关法规或应用要求的内容特征,为后续处理提供依据的服务。

3.6

数据流通分发　data circulation and distribution

按照数据流通需要,对数据进行脱敏、隐私化、标准化处理并对外输出开发。

3.7

数据流程服务工具平台　data process service tool platform

为 DPS 各项业务开发和提供用于服务操作、管理的工具软件及业务平台。

3.8

数据流程支持服务　data flow support service

支持 DPS 企业、从业者和相关机构提升能力、加强合作、便利交易的服务。

注:如业务培训、规范评价等。

3.9

数据需求方　data demand side

数据服务客户　data service customers

提出数据服务需求的机构。

注1:包括组织内部的部门和外部的机构,在本文件中统称为需求方。

注2:需求方一般包括行业用户、人工智能企业和行业应用开发企业和机构。

3.10

数据服务方　data service provider

数据服务供应商　data service provider

为需求方提供数据服务的机构。

注:包括组织内部的部门和外部的机构,在本文件中统称为服务方。

3.11

变更数据捕获　change data capture;CDC

主要用于捕获数据库的一些变更,然后可以把变更数据发送到下游的数据库领域的技术。

3.12

语音活动检测

从声音信号流里识别和消除长时间的静音期,以达到在不降低业务质量的情况下节省话路资源,是 IP 电话应用的重要组成部分。

注:又称语音端点检测和语音边界检测。

3.13

背景噪声　background noise

在发生、检查、测量或记录系统中与信号存在与否无关的一切干扰。

注1:在工业噪声或环境噪声测量中背景噪声指被测噪声源以外的周围环境噪声。如对在工厂附近的街道测量噪声来说,若要测量的是交通噪声,则工厂噪声便是背景噪声;若测量的目的在于测定工厂噪声,交通噪声便成为背景噪声。

注2:在噪声测量过程中,需要注意背景噪声的干扰程度。

注3:又称本底噪声底噪。

3.14

信噪比　signal to noise ratio

一个电子设备或者电子系统中信号与噪声的比例。

注1：信号指的是来自设备外部需要通过这台设备进行处理的电子信号，噪声是指经过该设备后产生的原信号中并不存在的无规则的额外信号(或信息)，并且该种信号并不随原信号的变化而变化。

注2：信噪比又称为讯噪比。

3.15

切音　segment

因录音操作导致的收音设备获取信号不完整，常表现为开始时或结尾处数据不完整。

3.16

截幅　cut off amplitude

把信号的幅值限制在某一固定的最大值的过程。

注：截幅也称为限幅。

3.17

采样率　sampling rate

每秒从连续信号中提取并组成离散信号的采样个数。

注1：采频率用赫兹(Hz)来表示。

注1：采频率也称为采样速度或者采样频率，采样频率的倒数是采样周期或者叫作采样时间，它是采样之间的时间间隔。通俗地讲采样频率是指计算机每秒钟采集多少个信号样本。

3.18

比特率　bit rate

每秒传送的比特(bit)数。

注：单位为bit/s，比特率越高，每秒传送数据就越多，音质就越清晰。声音中的比特率是指将模拟声音信号转换成数字声音信号后，单位时间内的二进制数据量，是间接衡量音频质量的一个指标。

3.19

声道　audio channel

声音在录制或播放时在不同空间位置采集或回放的相互独立的音频信号。

注：声道数是指声音录制时的音源数量或回放时相应的扬声器数量。声卡所支持的声道数是衡量声卡档次的重要指标之一，从单声道到最新的环绕立体声。

3.20

丢帧　frame loss

由设备引起的一段时间的信号丢失，即说话内容和底噪信息全部丢失。

3.21

前后预留　reserved before and after

一段音频实际说话段的前后非说话段。

3.22

噪声符号　noise symbol

代用以表示非说话段的，含有特殊意义的符号。

3.23

时间边界　time boundary

对一段语音数据在原始音频中的开始时间标记和结束时间标记。

3.24

转写　transliterate

将音频的内容由声音的形式转化为文字标记的过程或动作。

3.25

说话人　speaker

音频中发出声音的自然人。

注：包括以下类型：

a）　客服类说话人：客服类对话中的说话人，由客服人员和客户人员组成；

b）　访谈类说话人：访谈类对话中的说话人，由主持人和嘉宾组成；

c）　对话类说话人：自然对话中的说话人。

3.26

口音　oral speech sounds

一种对词或特殊音节的模块化强调程度的变化。这些变化都是由口腔肌肉和舌头的动作所产生。

注：口音是能够透过自小培养及练习而得来的。因此从口音能够反映人的出生地方或社会背景。学习某一种口音会使某个社会阶层产生认同感。演员学习口音为了使角色更传神。一个人的口音亦会随着居住地点的转移以及适应时间而产生变化。

3.27

语音合成　text to speech;TTS

将储存于电脑中的文件，如帮助文件或者网页，转换成自然语音输出的一种语音合成应用。

注：TTS不仅能帮助有视觉障碍的人阅读计算机上的信息，更能增加文本文档的可读性。现在的TTS应用包括语音驱动的邮件以及声音敏感系统，并常与声音识别程序一起使用。

4　总则

数据流程服务存在多种数据类型和服务场景，不同数据类型和不同服务场景的技术规范指标不完全相同，但对规范性的评价流程和指标统一。数据流程服务总体构成及其在数字产业中的功能如图1所示。

图1　数据流程服务总体构成及其在数字产业中的功能

数据流程服务技术采用自愿原则。

5　服务场景及其技术准则

由于不同业务场景对数据流程服务质量评价有不同的标准，因此本文件将数据流程服务技术分为通用技术和场景技术两方面。前者聚焦与场景无关的数据类型通用技术，后者聚焦于面向各业务场景的服务技术。

6 结构化数据流程服务技术

6.1 服务流程

6.1.1 概述

数据处理过程要有健全的流程管理机制，保障数据处理结果的准确性、完整性、一致性、时效性、可访问性和安全性，做到可溯源、可追踪。

结构化数据处理推荐流程见图 2。

图 2 结构化数据处理推荐流程

6.1.2 需求分析

根据需求方对数据处理的要求，明确采集数据源、采集范围、采集方式和频率，确定数据的合法合规和准确性，形成数据供需清单；编制信息资源目录，明确信息资源的分类、格式、数据项名称、数据类型、共享属性、开放属性、更新周期等内容，形成信息资源清单，并实现数据资源格式化和结构化。

6.1.3 数据采集

根据需求分析结论进行数据采集。

采集方式可采用程序内布码埋点、API 接口、网络爬虫、数据库对接、文件交换、邮件订阅等。

6.1.4 预处理

针对采集到的原始数据中存在的二义性、重复、不完整、违反业务规则等问题进行数据预处理。预处理任务可包括解析转换、纠错、异常值处理、缺失值处理、过滤、去重、标准化（格式标准化、值域标准化等）、数据入库等。

6.1.5 分析集成

分析集成是分析散落在不同数据源的业务实体数据，基于面向对象的数据组织原则，将数据按业务实体进行拉通并融合。高效的分析集成工作需要建立业务模型。

模型设计包括概念模型设计、逻辑模型设计、物理模型设计三个阶段。概念模型设计，用于识别核心业务流程，抽象业务流程中的实体和关系，定义业务域，完成实体关系的领域划分。逻辑模型设计，用于概念模型实体化，添加属性和属性定义，实体关系的梳理，归纳，定义物理化方式，添加必要的说明和描述。物理模型设计，需基于实际物理平台，完成平台的设置和优化，添加必要的元数据字段用于管理，生成最终建表语句并优化。

分析集成技术方法可包括数据挖掘、数据聚合、数据关联分析、聚类分析、假设检验等，属于大数据中关系挖掘的重要手段。

6.1.6 质量稽核

根据服务成果标准及稽核规则，对数据质量进行监控预警并产出质量报告。服务方要对数据质量负责。

数据质量按照 GB/T 36344—2018 中的指标说明，从数据的规范性、完整性、准确性、一致性、时效性和可访问性等多个层面实现对数据的全面稽核和预警，做到事前质量检查、事中运行监控、事后归纳总结，结合系统提供的全方位评估并提高数据质量，指导决策者的决定。

服务方应制定数据质量管理目标，建立相应的数据质量管理体系及实施机制、优化数据质量并持续改进，满足需求方数据应用的需求。对数据质量进行全程监控，做到数据质量全程追溯，可直接定位到问题数据所在的数据库→数据表→数据字段→数据值。且在监控到异常数据时应及时通过邮件、短信等方式通知到相关数据管理者。支持自动生成数据质量报告，以文字或图形化的方式展示数据质量及规范落地执行情况。支持数据质量规则的可视化配置，且实时展示数据质量规则的运行状态、运行结果。可保留数据质量规则的校验历史，数据质量的变化规律，找出数据问题。

6.1.7 流通分发

流通分发是根据需求方要求，将数据传输导入到目标系统中，实现数据在不同应用系统之间的共享或转移。

流通分发方式可包括数据资源目录、邮件订阅、API 接口、文件下载、离线文件交换。

6.2 服务技术

6.2.1 概述

完成流程中各环节服务任务所采用的技术及其规范。

服务技术评价指标一般情况下要根据业务场景确定标准值，同时不同水平的指标值也代表着不同水平的服务成本，推荐的评价指标值可参见本文件体系具体场景部分，如“第 4 部分：数据流程服务智慧园区场景技术规范”等。

6.2.2 数据采集技术

对数据采集进行安全管控，严格控制人员权限，采集数据和采集过程要有日志记录，保障数据采集可以追溯。保障采集传输安全和一致性，采用身份认证、数字签名、加密算法、SSL/TSL 传输协议等方式保障数据采集的安全性和完整性。

结构化数据采集技术功能和评价表见表 1。

表 1 结构化数据采集技术功能和评价表

序号	技术名称	功能描述	评价指标或原则
1	数据库直连	基于现有数据库，通过指定路径快速安全访问并获取数据。通过 JDBC 方式直连业务系统，根据约定周期抽取数据	增量数据的更新频率、访问和获取数据的便利性
2	CDC	通过解析日志变化情况，获取数据变更	数据延迟、数据库负载影响、数据变更状态完整性
3	SDK	业务软件集成 SDK；通过 SDK 直连抓取数据	数据延迟、采集数据完整性、更新难度
4	API 接口	业务系统通过调用应用程序的接口实时采集数据，基于不同数据接口按需调取，快速对接自己的数据库	接口可用率。接口可用率＝(服务可用时间/服务总时间)×100%； 接口查询延迟时间

表 1　结构化数据采集技术功能和评价表（续）

序号	技术名称	功能描述	评价指标或原则
5	物联网设备直传	根据物联网设备通信传输协议直接接入按需调取，快速对接自己的数据库	物联网协议覆盖度、并发数、接口延迟时间
6	网络爬虫	主动抓取互联网上所需数据，实现全网内容批量更新和重点信息实时更新	有效性（对抗验证码、防火墙策略）、数据时效性

6.2.3　预处理技术

结构化数据预处理技术功能和评价表见表 2。

表 2　结构化数据预处理技术功能和评价表

序号	技术名称	功能描述	评价指标或原则
1	解析转换	信息实时解析，将非结构化、半结构化数据转换成为结构化数据	服务成本变化率、数据时效性、数据维度可扩展性
2	纠错	利用统计分析或人工智能的方法检测属性可能的错误值或异常值，并加以修正。按照属性值的平均值或中值来替换属性值；简单规则库（常识性规则和业务特定规则等）检测和修正错误；使用不同属性间的约束检测和修正错误；使用外部数据源检测和修正错误	规则库完整性、自动化程度
3	异常值处理	异常值数据指无意义的数据、坏数据，包含所有难以被机器正确理解和翻译的数据，如非结构化文本。删除含有异常值的记录；将异常值视为缺失值，按照缺失值进行处理；可用前后两个值的平均值修正；不直接在具有异常值的数据集上进行数据挖掘	规则库完整性、自动化程度
4	缺失值处理	指值实际存在，但没有存入值所属字段。可以从本数据源或其他数据源推导出来；可用平均值、中间值、最大值、最小值或更为复杂的概率统计函数值代替缺失的值，但准确性比较低；人工输入一个可接受的值	规则库完整性、自动化程度
5	空值处理	完整性检查，检查表中某一列字段数据是否含有空值	规则库完整性、自动化程度
6	去重	检查数据是否唯一，识别表中的重复数据	规则库完整性、自动化程度
7	格式标准化	基于数据元标准检查表中某字段数据的数据格式是否正确	根据业务内涵订立标准。例如：邮箱格式、身份证格式等
8	值域标准化	检查表中某字段数据取值是否在指定范围或指定维度值内	根据业务内涵订立标准。参照相关领域国标或行标。例如：年龄字段数据取值范围是否在0～150内
9	记录数	检查一张表记录条数是否在指定阈值范围内或与历史数据比较波动值是否在一定范围内	根据业务内涵订立标准。例如：检查“用户访问明细表”今日新增记录数与昨日相比上下波动范围是否在－10%～10%内

表 2　结构化数据预处理技术功能和评价表（续）

序号	技术名称	功能描述	评价指标或原则
10	逻辑性	对表内或两张表间的某一列数据或某几列数据的表达式与其他某一列或某几列数据的表达式比较，检查数据逻辑是否正确	例如：对“可视电话用根据业务内涵订立标准。户情况统计表”逻辑检验：未超出套餐使用量的活跃用户数＋超出套餐使用量的活跃用户数＝活跃用户数
11	及时性	检查单表数据更新时间是否在指定时间范围内	根据需求订立标准
12	拉链表检查	检查拉链表的数据是否有断链、交叉链、重复链	根据业务内涵订立断链率、交叉率、重复率标准
13	自定义	校验数据是否符合用户自定义SQL脚本内容	根据需求订立标准

6.2.4　分析集成技术

结构化数据分析集成技术功能和评价表见表 3。

表 3　结构化数据分析集成技术功能和评价表

序号	技术名称	功能描述	评价指标或原则
1	关联数据存储	将实体、关系按照时间组成时空网络，实现在各时间粒度“回放”事物的发展过程，从微观角度看清来龙去脉；从宏观角度看到同类事物的共同发展通性，对未来进行预测，是图存储和图分析技术的统一体	存储有效性和基于图的关联查询时长
2	关联数据快速分析	关联分析算法基于多种存储方案，包括以Oracle、MySQL为代表的传统型关系数据库，以Hadoop、HBase为代表的键值对的存储方案	数据分析结果返回时长
3	关联数据和传统数据同步更新	该系统解决方案既包含宏观大数据、更突出微观大数据的统一体	处理完备性(增、删、改)； 关系数据库实时性、同步率
4	聚类分析	聚类分析用于洞察数据的分布，获取数据的特征和进行异常检测	分类主题合理性、分析耗时
5	假设检验	利用抽取的样本信息去判断总体假设是否合理，即判断总体的真实情况与假设是否存在显著的系统性差异	差异检测准确度、分析耗时

6.2.5　质量稽核技术

结构化数据质量稽核技术功能和评价规范表见表 4。

表 4　结构化数据质量稽核技术功能和评价规范表

序号	技术名称	功能描述	评价指标或原则
1	数据稽核	可通过全表扫描或者数据抽样方式对数据进行检查，数据抽样分情况使用分层抽样和随机抽样方式对数据进行抽查	数据抽样应具备代表性

表 4　结构化数据质量稽核技术功能和评价规范表（续）

序号	技术名称	功能描述	评价指标或原则
2	稽核规则	应支持值阈检查、规范检查、逻辑检查、及时性检查、完整性检查、波动性检查和自定义SQL检查	规则支持度。如果SQL结果或以上检查不在值阈范围，则触发报警
3	质量稽核报告	应体现数据完整性、规范性、一致性、准确性、关联性，能通过报告及时并快速定位问题	报告完整性。数据来源、指标定义、数据处理、报告结论
4	数据波动性	检查表中某字段数据值对比之前业务周期数据值的浮动是否在一定范围内	根据业务内涵订立标准。例如：校验“商品收益表”中某商品今日收益总额与昨日相比上下波动范围是否在－5%～5%内

6.2.6　流通分发技术

流通分发方式分为服务方主动推送和需求方主动拉取，需求方可通过数据订阅设置推送方式和时间来完成数据获取，也可通过申请秘钥通过 API 接口或文件方式主动拉取完成数据获取。需求方依靠数据资源目录和元数据来管理、解读获取的数据。

结构化数据流通分发技术功能和评价表见表 5。

表 5　结构化数据流通分发技术功能和评价表

序号	技术名称	功能描述	评价指标或原则
1	数据资源目录	提供统一的数据资源视图，为数据生产者、管理者、使用者提供快速查询入口。资源编目可按四个角度对数据服务进行分类，包括：组织机构、业务主题、管理专题三个维度。包括数据目录及服务目录，第三方服务可以直接挂载，然后通过服务目录对外提供服务。数据目录支持API接口、文件、数据库交换等方式。支持数据需求者通过数据资源目录进行数据访问申请，数据管理者进行访问授权，通过的用户可以通过接口、数据库、文件等多种方式进行数据使用	分类合理性、存储效率、检索时长
2	元数据管理	元数据是对数据的描述，通过描述数据的模型、产生、使用、业务含义、数据所有者等信息，帮助数据使用方了解和使用数据。元数据分为业务元数据、技术元数据和管理元数据。业务元数据描述数据的来源、数据字典、业务含义、统计数据；技术元数据描述数据的存储情况、血缘关系、质量稽核报告；管理元数据描述数据的所有者、权限使用范围、分类分级、数据冷热情况	数据找得到、数据读得懂，数据语言统一。采集和管理范围、应用情况和范围、标准化程度、自动化程度

表 5　结构化数据流通分发技术功能和评价表（续）

序号	技术名称	功能描述	评价指标或原则
3	API接口	基于云端百种数据接口，企业可按需调取，快速搭建自己的数据库，满足企业实现低成本、高效调用数据的需求	按需最小化原则，保障数据传输的安全性，保障数据服务合法、可控、可追溯和权责一致。 服务稳定性、服务性能、数据传输的安全性
4	在线文件下载	通过在线按需查找相关数据，将查询结果传送到本地计算机磁盘上并保存起来	
5	离线文件交换	将本地存储的文件上传到目标数据平台，按照指定的方式获取所需要的结果。 通过FTP、SFTP等文件传输协议进行离线文件传输，完成数据交换	
6	邮件订阅	通过对多维度数据动态抓取和监控预警，以邮件形式把数据的动态变化发送到用户邮箱中，系统性地解决对企业的全面尽调和监控中的痛点。 在线进行数据订阅，并通过邮件发送和接收	
7	数据联邦	通过访问一个全局虚拟数据库，通过全局虚拟数据库管理系统将分布在不同物理数据库中的数据抽象成一个统一的数据视图，为不同的应用系统提供全局信息服务，实现不同应用系统和数据源之间的信息共享和数据交换。数据联邦实施应考虑数据安全、数据延时、数据的有效性、数据的一致性和质量，以及数据的可用性、数据模型改变的影响、性能、数据访问量等一系列问题	模型的效果（评价指标KS/AUC等）、联邦学习耗时、安全性（参照GB/T 37988）
注：AUC值：衡量模型对于正负样本的整体区分能力。 KS值：衡量模型对于正负样本的最佳区分情况，区分度越大说明模型的风险排序能力越强，与AUC结合使用判断。			

6.2.7　数据安全技术

数据安全：数据安全应构建数据的分级分类机制，建立数据应用、管理、备份和恢复的安全保护管理机制和策略，对数据完整性、保密性、隐私性、可信性等进行保护。数据安全涉及阶段包括：采集、预处理、分析集成和流通分发 4 个阶段。

结构化数据安全技术功能和评价表见表 6。

表 6　结构化数据安全技术功能和评价表

序号	技术名称	功能描述	评价指标或原则
1	数据分类分级	重要数据、核心数据和个人信息、其他分级数据中敏感数据要进行数据加密脱敏，避免在存储还传输过程泄露或出现数据越权导致的数据安全问题	参考《网络安全标准实践指南——网络数据分类分级指引》
2	安全审计	建立安全审计规章制度和管理机制；建设安全审计组织团队；结合大数据和人工智能技术建设安全审计平台；定期开展安全审计工作，防范于未然	审计完整性[a]、管理和预防能力、技术先进性
3	身份认证	鉴别通信中另一端的真实身份，防止伪造和假冒等情况发生，包括数字签名、数字证书、匿名认证	完整性、真实性、不可否认性

表 6　结构化数据安全技术功能和评价表（续）

序号	技术名称	功能描述	评价指标或原则
4	访问控制	数据库库表级、行列权限、接口IP白名单、TPS限制	控制粒度
5	数据加密、脱敏	对数据加密脱敏，防止数据主观、不经意或被动泄露	运算速度、安全性、资源消耗
6	数据区块链	对数据资产进行确权，授权和鉴权，并且调动数据计算引擎，实现数据用途和用量的可控	满足央行、工信部等主管部门的相关测评要求
7	隐私计算	在保证数据提供方不泄露原始数据的前提下，对数据进行分析计算的一类信息技术，保障数据在流通和融合过程中的各个环节中“可用不可见”	保证输入数据保密性是整个隐私计算过程中不泄露数据提供方的原始数据
[a] 审计完整性包括： a）　审计范围应覆盖到服务器和重要客户端上的每个操作系统用户和数据库用户； b）　审计内容应包括重要用户行为、系统资源的异常使用和重要系统命令的使用等系统内重要的安全相关事件； c）　审计记录应包括事件的日期、时间、类型、主体标识、客体标识和结果等； d）　应保护审计记录，避免受到未预期的删除、修改或覆盖等； e）　应能够根据记录数据进行分析，并生成审计报表； f）　应保护审计进程，避免受到未预期的中断。			

6.3　服务成果

数据流程服务成果即为加工处理后的数据，其标准保障数据内外部使用一致性和准确性，解决数据指标中同名不同径，同径不同名，口径不清晰，命名难理解，计算不易懂等问题，提升数据治理、加快数据流通、避免数据歧义。

数据标准分为业务数据标准和技术数据标准。业务数据标准分为主数据标准、元数据标准、指标标准；技术数据标准描述了数据模型定义规范，数据模型应有统一的命名规范、数据类型，以便于数据理解和流通。

主数据标准参照各业务场景相关领域标准，如《世界各国和地区名称代码》《表示货币和资金的代码》《术语工作》系列标准。

元数据标准，影响数据流通，相关数据标准参照 GB/T 7408、GB/T 18391(所有部分)等。

技术数据标准参照各行业相关标准，如 GB/T 34077.1、GB/T 33767(所有部分)等。

7　非结构化数据流程服务技术

7.1　语音数据服务技术

7.1.1　语音数据采集和预处理技术指标及测量

7.1.1.1　采集设定流程

数据采集是指按照指定需求场景要求，使用指定设备，收集并交付原始数据。采集流程首要需要明确采集对象、设备要求，通过试采环节确定交付数据标准，并在指定时间内完成数据交付。采集设定流程如图 3 所示。

图 3　采集设定流程

a）　需求确定阶段：

——本阶段的目标是解读并充分理解需求，明确合格标准，确定需求内容无歧义；

——本阶段的主要任务包括确定采集对象，采集工具及参数；

——采集对象确定内容包括但不限于采集目标类型、采集目标数量、采集目标分布、采集目标环境、采集目标形态等；

——采集工具及参数确定内容包括但不限于采集工具类型、采集工具分布、采集工具搭设、采集工具调参、采集工具配合、采集工艺等。

b）　试采阶段：

——本阶段的目标是为正式采集进行前期准备，通过试采部分数据，验证服务成果是否符合需求确定内容；

——本阶段的主要任务包括试采集，试采集质量检验和判定标准确定，输出数据采集方案；

——试采集：按照采集文档要求产出第一批数据，通过内外规则及软硬件磨合，反哺采集工艺及采集流程，保障采集工艺可复现，采集流程可实施；

——试采集质量检验和判定标准确定：与需求方、项目管理者、项目执行团队确定质量验收标准，并对试采集数据进行检验。质量验收标准参见 7.1.1.3；

——输出数据采集整体方案包括数据采集方案、采集质量控制方案、采集风险预案、项目执行日志、数据存储方案等。

c）　正式采集阶段：

——本阶段的目标是完成规定任务量的采集服务；

——本阶段的主要任务是按照数据采集整体方案，保障采集工期、采集计划顺利执行，达到采集质量和交付要求。

d）　质检阶段：

——本阶段的目标是保障正式采集的数据符合质量验收标准；

——本阶段的主要任务是对采集好的数据进行清洗和质检，审核方式可以采用同步审核和完成抽检，并且根据检验情况可与正采阶段形成多轮循环流程；

——同步审核：对采集数据实时质量监控，保障采集质量控制方案落地，采集过程及时纠偏，输出阶段性采集数据指标；

——完成抽检：对质检数据进行比例抽检，验证同步审核结果准确且置信，数据成果符合质量验收标准。

e）　交付阶段：

——本阶段的目标是完成项目，赢得需求方满意；

——本阶段的主要任务是将符合质量验收标准的合格数据交付，包括数据验收、数据结算和需求交付；

——数据验收：需求方对提交的全量数据进行比例抽检，确认数据可用且符合制定的质量验收标准要求，满足则触发数据合格结算，不满足预期则进行纠偏返修；

——数据结算：对交付的符合验收指标的有效数据进行结算信息确认，包括但不限于报价信息、关键指标、数据量级、结算账期等；

——需求交付：实现流程闭环，输出采集交付报告，原始数据及授权文件回传，设备返还，调研服务满意度等。

7.1.1.2 采集工具要求

语音采集工具通常包括采集设备和软件平台。

采集设备有麦克风、麦克风阵列、手机、录音笔、专业录音棚及工业级录音设备。手机、录音笔等，用于常规语音数据的采集；专业录音设备用于高标准数据的采集。采集设备功能要求符合相关设备质量标准。

采集软件平台功能要求一般包括：

a） 可以实现批量人员同时并行录制采集；

b） 采集文本可以展示在独特的 UI 界面上，方便用户注意信息文本，保证语音采集要求；

c） 结合云服务开发，可降低数据传输风险；

d） 过程需有管理逻辑，用户管理、文本管理和声音文件管理齐全，方便管理员进行审核提交；

e） 语音采集 APP 支持手机操作习惯，方便暂停或重启后继续录制，能避免用户漏录和错录；

f） 音频数据展示，方便了解语音详细数据，指导正确语音采集。

7.1.1.3 采集质量标准

采集质量涉及采集服务成果质量和流程质量两方面。

语音数据采集流程质量要素及测量方式见表 7。

表 7 语音数据采集流程质量要素及测量方式

指标名称	指标定义及要求	计算逻辑
一次交付达成率	项目数据一次交付时准确率的达成情况，用于衡量项目的质量保证能力	一次交付达成率＝(一次交付准确率/目标准确率)×100%
单次交付合格率	项目数据每次(按项目与业务约定的交付周期：日、周、月、数据包)交付时准确率的达成情况，用于评估交付能力	单次交付合格率＝1－(项目单次交付不合格数量/项目单次交付数量)×100%
终审交付达成率	项目数据终审通过的数量，用于衡量项目的交付质量情况	终审交付达成率＝(终审交付的合格数据量/目标交付合格的数据量)×100%
交付周期延时率	量级要求项目：实际交付周期与目标交付周期的时间差。用于评估交付能力	交付延时率＝[(实际交付周期－约定交付周期)/约定交付周期]×100%

语音采集项目类型一般有唤醒词采集、命令词采集、普通文本朗读采集、自然对话文本采集、会议数据采集、其他噪声采集等。每一个子类项目各自质量检验侧重点不同，如无特殊要求，可参照正确性检验规范来要求。正确性检验规范应包含数据的采样率、比特率、声道、前后预留、切音、截幅、丢帧、底噪、混响、响度、信噪比、口音要求等相关的指标量化要求。

7.1.1.4 采集数据格式

采集格式：一般为 wav，mp3，v3，m4a，pcm 等格式音频文件。

预处理格式：json、xml、txt 等格式。

采集数据的包装格式推荐按照以下格式：

——xx 语音数据库；

——wav；

——speakerid；

——00001.wav；

——userinfo.txt。

音频文件存储格式为，总文件夹名称 wav，子文件夹名称为说话人编号，同一个说话人的语音在一个子文件夹中。

Userinfo 为录音人的说话人信息，考虑到个人隐私和数据合规性，采集数据需要获取采集人的授权协议，同时记录的信息遵循最小化原则，性别、年龄、地域籍贯，若项目对设备机型或距离有要求的，也需记录，如图 4 所示。

录音人ID	录音日期	录音地点	录音设备	性别	年龄	籍贯
A001	2018/11/13	北京	huawei P6	女	24	重庆

图 4　采集数据记录示例

7.1.2　语音数据标注/内容审核技术指标及测量

7.1.2.1　标注/审核设定流程

标注需求承接需要将各种类型数据进行集成封装，通过数据载体或平台，传输到标注侧进行人工分类识别处理，标注侧最终对需求侧提供标准化的、可供检索、分析或可视化的数据分类服务交付。标注服务流程如图 5 所示。

图 5　标注服务流程

a）　需求确定阶段：

——本阶段的目标是解读并充分理解需求，明确合格标准，确定需求内容无歧义；

——本阶段的主要任务包括确定需求对接，需求评估；

——需求对接：了解标注背景、数据源特征、数据密级、交付工期、验收指标、作业要求等；

——需求评估：分析需求侧强关注指标及需求成果，规划资源。

b）　试标阶段：

——本阶段的目标是为正式标注/审核进行前期准备，通过试标部分数据，验证服务成果是否符合需求确定内容；

——本阶段的主要任务包括需求承接，试标质量检验和判定标准确定，制定标注/审核方案；

——需求承接：在安全密级、成本、资源的综合平衡下，选择匹配的承接团队进行试标，并在试标过程中进行标准优化、标注模式确认、人效测试、成本评估等；

——试标质量检验和判定标准确定，与需求方、项目管理者、项目执行团队确定质量验收标准，并对试标注/审核数据进行检验。质量验收标准参见 7.1.2.3；

——制定方案：制定数据解决方案，规划进度、质量、成本管控细节，确认报价、工期、验收流程等关键信息，对齐需求侧，最终落地标注管理预案。

c）　正式标注/审核阶段：

——本阶段的目标是完成规定任务量的标注/审核服务；

——本阶段的主要任务进行标注管理，即观察数据源是否符合分类标准使用需求，对标注周

期、质量进行跟进，标注突发风险识别及处理。

d） 质检阶段：

——本阶段的目标是保障正式标注/审核的数据符合质量验收标准；

——本阶段的主要任务是对标注好的数据进行质检并且根据检验情况可与正式标注/审核阶段形成多轮循环流程。

e） 交付阶段：

——本阶段的目标是完成项目，赢得需求方满意；

——本阶段的主要任务是将符合质量验收标准的合格数据交付，包括数据验收、数据结算和需求交付；

——数据验收：测算质检结果合格率，与需求侧确认数据处理各项指标是否符合需求侧预期指标，符合验收要求则对需求侧验收结论书面输出，交付标注结果；

——数据结算：对交付的符合验收指标的有效数据进行结算信息确认，包括但不限于报价信息、关键指标、数据量级、结算账期等；

——需求交付：实现流程闭环，对需求侧提供完整的交付报告、调研满意程度、持续提供售后服务。

7.1.2.2 标注/审核工具要求

智能语音应用的实现涉及多种语音处理技术，如 ASR（语音识别）、NLP（自然语言处理）、TTS（语音合成）、Wake up（语音唤醒）、Voice Print（声纹识别）、DM（对话管理）等。其中最为重要且应用广泛的，主要有 ASR、NLP、TTS：

——ASR：语音转文本，相当于是该智能系统中的“耳朵”；

——NLP：自然语言理解，对文本信息进行处理，并做出对应指令，相当于是该智能系统中的“大脑”；

——TTS：文本转语音，相当于是该智能系统当中的“嘴巴”。

其中 NLP 属于文本数据服务技术，参见 7.2。语音数据流程服务主要涉及 ASR、TTS 两种技术。

对于语音标注工具的要求，主要有：

——能够实现音频截取、音频属性判断、音频文本转写等基本操作；

——含波形图，有明晰的时刻刻度，标注页面能够对波形图进行缩放；

——能够便捷的对音频进行播放、暂停、重新播放、调整播放速度、音频快进快退功能，并有对应快捷键辅助。

对于 ASR 审核工具的要求，除了满足上述标注工具的要求外，还需：

——对于发现标注数据所存在的错误，能够进行标识及校对；

——能够对错误数据进行驳回，打回至标注环节修改。

此外，标注/审核工具，还需要具备有方便、快捷、可视化的数据流转和统计的功能，便于利用数据化进行标注作业流程管理。

7.1.2.3 标注/审核质量标准

7.1.2.3.1 ASR

在语音 ASR 转写当中，主要操作为对音频进行截取、对截取部分音频进行文本转写、对截取部分音频进行属性判断，各环节操作规范如下：

——音频截取，根据实际应用场景需要对音频进行分割截取，并保证所截取音频与理想分段音频贴合；

——音频文本转写，将音频文件内容用汉字表示转写为文本，转写内容需要和实际发音内容完全

一致，不准许出现修改和删减的问题；

——音频属性判断，确定是否包含有效语音，确定语音的噪声情况，确定说话人数量，确定说话人性别，确定是否包括口音。

其中对于标注是否符合规范的判断标准和依据，如表 8 所示。

表 8　语音数据标注/审核 ASR 流程质量要素及测量方式

操作	技术规范	规范逻辑	规范说明和要求
音频截取	音频截取贴合程度	—	实际截取音频与理想截取音频尽可能贴合，不过多截取导致音频缺失，也不留白过多，具体指标值需根据业务场景确定
音频文本转写	字准率/句准率	句准率=1−句错率； 句错率(SER)=(错误句数/总句数)×100% 字准率=1−字错率； 字错率(WER)=(错误字数/总字数)×100%	字错率是语音识别领域的关键性评估指标，WER越低表示效果越好；根据应用场景不同以及语音检测工具不同，对于高质量音频转写要求，要求有所不同。具体指标值需根据业务场景确定
音频属性判断	音频属性判断合格率	音频属性判断合格率=(音频属性判断正确数/音频属性判断总数)×100%	根据应用场景不同以及语音检测工具不同，对于高质量音频判断要求。具体指标值需根据业务场景确定

7.1.2.3.2　TTS

在 TTS 语音合成当中，主要的标注任务包括：文本语料收集、文本对齐、断句切分、拼音、韵律精标、音素切分、主观评测、离线测评等。

同时，任务涉及“全局区间”“局部区间”“帧”的标注层设置。

——全局区间：针对整条音频进行标注，全局区间的起止位置即整条音频的开始结束位置，主要是标注一些全局信息，如整条音频的转写内容，语种等。

——局部区间：针对切分出的部分音频段进行标注，主要标注拼音、音素以及其他针对部分时间段标注的信息。

——帧：指标注音频的某一时刻，主要标注韵律。

语音数据标注/审核 TTS 流程质量要素及测量方式见表 9。

表 9　语音数据标注/审核 TTS 流程质量要素及测量方式

操作	操作定义	技术规范
文本语料收集	通过音、视频等材料收集文本并进行顺滑整理，为录音提供素材	——依据规则文档把握制作音库的需求特征； ——寻找错别字少、标点规范、三观正常的内容
文本对齐	对音频和文本内容进行校对处理	——每个音频里发音人字音朗读准确，无音频质量问题、无发音准确性问题； ——文本内容和音频内容一致； ——文本内容是规范和正则化的中文表达； ——标点符号使用规范

表 9　语音数据标注/审核 TTS 流程质量要素及测量方式（续）

操作	操作定义	技术规范
断句切分	对长片段语音数据进行断句处理，得到音频与对应的规范后的文本	——切出的语句最好要语法正确，语义完整，语气完整，无音频质量问题、无发音准确性问题； ——文本内容是规范和正则化的中文表达； ——标点符号使用规范
拼音精标	对预测后文件进行检查	——根据读音进行拼音的声调和音素拼写标注； ——判断是否有停顿，标注出停顿的相对位置，停顿的顺序位置也要准确； ——保证第一层文本内容与第二层汉字内容完全一致
韵律精标	对预测后文件进行检查	——对音节、音步、韵律词、韵律短语、语调短语、句子等韵律等级进行划分； ——保证第一层文本内容与第二层汉字内容完全一致
音素切分	做完拼音检查后，将语音按照给定音素序列进行强制切分，得到每一个音素的时间段信息	——根据标注规范调整音素边界； ——根据停顿情况增加或修改停顿边界
主观评测（mos）	对于给定的语音，试听完根据第一感受，给出主观评分（MOS-Mean Opinion Score，即平均主观意见分）	——从不同维度对单条数据的整体感受打分； ——根据第一语感进行打分
离线测评	对于给定的语音&文本，先听语音，再根据试听结果，结合文本比对，找出语音片段中的前后端错误	区分不同音库需要反馈的前后端问题，前端类比人类的语言中枢（根据句子预测应该读音和停顿），后端类比人类的发音器官（根据预测结果发声）

7.1.2.4　标注/审核数据格式

7.1.2.4.1　ASR

音频文件：一般为 wav 音频格式，包含原始音频及切分后子音频；

音频标注文件：标注文件格式常见的有 json、txt、TextGrid、csv 等，包含内容：

——被截取音频片段的起止时间戳；

——该截取音频的有关属性，如有效无效、噪声、说话人性别等；

——该截取音频对应的转写文本等。

7.1.2.4.2　TTS

音频文件：一般为长短不一的 wav 音频格式；

音频标注文件：标注文件格式常见的有 pychk、prsdchk、TextGrid、intervial、csv 等。

7.1.3　语音数据流通分发技术指标及测量

7.1.3.1　流通分发设定流程

数据分发是根据数据应用要求，将数据服务成果交付需求方。需求方包含公司内外部客户，交付数据包含客户定制项目，也包括自有数据。数据流通分发设定流程如图 6 所示。

图 6　数据流通分发设定流程

a）　需求确定阶段：根据业务需求评估流通数据类型、体量、要素等，约定数据传输方案，与内部服务团队确定数据资源，将获取的少量数据样本与承接方进行共享，基于样本评估需求落地可行性。

b）　协议确定阶段：进行合规评估，识别敏感数据所属类别及敏感级别，确保风险可控；签订保密协议。

c）　流通处理阶段：

——数据获取：对获得采集许可的数据进行批量入库，实际获取的数量、来源等对齐评估部门；

——数据脱敏：对于涉及用户隐私的数据，采用信息加密、信息替换、信息模糊化等策略和技术方法进行数据脱敏。

d）　交付阶段：按约定方式回传数据及结果，或采用隐私计算技术提供服务。通过本地上传、API、公司数据库流转等渠道将数据分发至指定需求方。

7.1.3.2　流通分发工具要求

流通分发可采用多种方式和工具进行：

——API 接口：通过 API 接口进行数据传输；

——Url 访问：通过 url 链接对数据进行访问；

——FTP 传输：使用 ftp 传输工具进行数据传输；

——私有云传输：通过密钥访问私有云数据；

——其他工具：网盘、硬盘等工具。

7.1.3.3　流通分发质量标准

流通分发质量由内容质量和传输质量构成。数据内容质量参照 7.1.1.3 采集质量标准和 7.1.2.3 标注/审核质量标准。传输质量参照相关传输方式技术规范，其中 API 接口标准参照以下内容。

a）　API 接口定义：可通过平台预先定义的函数或一种约定协议，对平台或工具发起数据服务请求。如上传数据、下载数据。

b）　规范要求：

——API 与客户端用户的通信协议，尽量使用 HTTPS 协议，以确保交互数据的传输安全；

——宜尽量将 API 部署在专用域名之下，如果确定 API 很简单，不会有进一步扩展，可以放在主域名下。

c）　常用场景：

——API 数据送标、API 数据结果导出；

——API 数据送标：指平台提供 appId、appKey 等关键字段信息，其他平台或工具可通过访问请求，上传给标注平台相关数据；

——API 数据结果导出：指平台提供 appId、appKey 等关键字段信息，其他平台或工具可通过访问请求的方式，获取及下周标注平台的相关数据。

7.1.3.4 流通分发数据格式

流通分发数据格式参照 7.1.1.4 采集数据格式和 7.1.2.4 标注/审核数据格式。

7.2 文本数据服务技术

7.2.1 文本数据采集和预处理技术指标及测量

7.2.1.1 采集设定流程

参照 7.1.1.1

7.2.1.2 采集工具要求

文本采集涉及两类场景，一类针对新闻资讯类、行业互联网和政府开放的数据，就是将非结构化的网络文本信息从大量的网页中抽取出来保存到结构化的数据库中的过程。收集后的海量内容素材，经过数据清洗过程，处理成可用于标注的文本内容。另一类针对某指定语义内容进行泛化，文本采集途径通常为人工采集，可人工使用 EXCEL 汇总或某些数据服务平台工具编制。

7.2.1.3 采集质量标准

采集质量涉及采集服务成果质量和流程质量两方面。

文本数据采集流程质量要素及测量方式见表 10。

表 10 文本数据采集流程质量要素及测量方式

指标名称	指标定义及要求	计算逻辑
一次交付达成率	项目数据一次交付时准确率的达成情况，用于衡量项目的质量保证能力	一次交付达成率＝一次交付准确率/目标准确率×100%
单次交付合格率	项目数据每次（按项目与业务约定的交付周期：日、周、月、数据包）交付时准确率的达成情况，用于评估交付能力	单次交付合格率＝1－（项目单次交付不合格数量/项目单次交付数量）
终审交付达成率	项目数据终审通过的数量，用于衡量项目的交付质量情况	终审交付达成率＝终审交付的合格数据量/目标交付合格的数据量×100%
交付周期延时率	量级要求项目：实际交付周期与目标交付周期的时间差。用于评估交付能力	交付延时率＝（实际交付周期－约定交付周期）/约定交付周期

文本采集项目类型一般分为线上采集、线下采集，线上采集例如评论采集、留言信息采集、文章采集等，线下采集例如语句扩写、对话采集等，每个采集任务的质量质检侧重点根据项目具体需求会有所不同，如无特殊要求，可参照正确性质检规范来要求。正确性检验规范包含数据的准确性、相关性、逻辑正确性、常识正确性、合规合法等相关指标量化要求。

a） 待标注数据——不带预标注结果：

——csv 第 1 列存放文本内容，列名不能为空，可自定义（中英文均可）；

——第 2 列列名不能为空，可自定义（中英文均可）。

data	result
＜文本内容＞	

b） 待标注数据——带预标注结果：

——csv 第 1 列存放文本内容，列名不能为空，可自定义（中英文均可）；

——第 2 列存放 json 结构的预标注结果，列名不能为空，可自定义（中英文均可）。

data	result
＜文本内容＞	

{"item":{"src":" 三、劳动报酬\n 1．甲方原意执行国家有关工资支付的规定，遵守最低工资保障制度。甲方工资发放日为每月____日；工资发放形式为直接发放 / 委托第三方代发。","templateID":"[redacted]type":"text-split","data":[{"type":"text","data":[{"type":"single-select","required":true,"description":"全文属性","options":["1","2","3"],"result

7.2.1.4 采集数据格式

文字数据处理支持 csv、url、doc 等常用交付格式，内含标注字段如 UID、垂类领域、关键词、标题、评论、前端原文链接等。

7.2.2 文本数据标注/内容审核技术指标及测量

7.2.2.1 标注/审核设定流程

参照 7.1.2.1。

7.2.2.2 标注/审核工具要求

文本数据标注/审核工具平台功能要求见表 11。

表 11 文本数据标注/审核工具平台功能要求

文本处理类型	模板元素	平台功能要求	适用项目类型
多模态通用	下拉框	多级分类打标签，常用于行业标签体系建设类标注需求，可以在各级子类目下按层次选择，使内容分类唯一性、精确化	文本分类/打分排序
	单选标签	适用于从多标签中选择唯一正确的标注答案的场景，可以提升标注答案精确性，提升标注结果判断的客观性	文本判断/文本评测/打分排序/文本审核
	复选框	复选框是从一组选项中选择一项或多选项的组件，可以满足一个 case 有多个答案的需求，具有高灵活性、容错率	文本判断/文本评测/文本审核
	文本输入框	文本输入框是页面中录入信息的组件，通常有两大作用。一是通过基础字段，在文本输入框分析关键信息，如：对原文文本进行信息提取/摘要/转写/扩充/翻译/清洗/优化等；二是出现在含有表单和对话框的标注需求中，可以备注、补充录入信息，使标注结果更精确完善	文本清洗/文本修改/文本翻译/文本摘要/提取/文本转写/扩充/词条优化
NER	划词标记	为固定文本选中标记内容，方便对关键文本进行特殊备注	文本摘要/提取/词条优化/文本审核

文本数据标注/审核辅助工具说明见表 12。

表 12　文本数据标注/审核辅助工具说明

辅助工具	说明
划词插件	将需要修改的词进行滑动选中，修改；对于长篇幅内容标签的标注，需要快速划词定位标签信息，可以节约大部分浏览成本
高亮插件	对于文字型的项目，能够直接高亮关键词定位，二次选择时不保留第一次定位结果
在线词典	专业划词翻译插件，依托大量权威词典涵盖中英日韩法德西语的交叉翻译功能，能够简单实现查词和翻译的功能
划词小窗搜索	可以通过小窗口打开搜索工具进行搜索，缩短打开网页以及复制粘贴的时间，进而提高人效
字数统计	适用于所有网页，可以实时自动计算选中的字数，减少字数占比统计的人工用时
划词翻译	划词即显示翻译结果，支持多有道翻译、百度翻译、谷歌翻译等多种翻译引擎搜索，支持多种语言的朗读，能翻译 PDF 文档
增强复制	可以复制标签页全部内容/多标签页，支持划词复制，减少了频繁 ctrl＋C、ctrl＋V 的操作时间

标注功能包括文本分类、实体词抽取（即文本切分）、实体关系标注：

——文本分类：可对一段文字内容进行分类标注或文字转写；

——实体词抽取：可对一段文字中某句话或某个词汇进行分类标注或文字转写；

——实体关系标注：可对两个实体词抽取结果，关联二者之间的关系；

——一键标注相同文本：手动选择一段文本内容后，会同时将全文中其他相同内容且未标注的内容也切分为文本段，并且使得它们属性值保持一致，直至切换选中其他文本段。

示例：

——全文存在 10 个“您好”；

——将第 1 个“您好”新增为文本段且标签选为“标签 A”，第 2 个～10 个“您好”自动新增为文本段且标签也为“您好”。

7.2.2.3　标注/审核质量标准

文本数据标注/审核流程质量要素及测量方式见表 13。

表 13　文本数据标注/审核流程质量要素及测量方式

适用项目类型	指标名称	指标定义及要求	计算逻辑
文本分类	正负例准确率	即一级分类准确率，“属于当前一级分类”为正例，“不属于当前一级分类”为负例。正负例准确率是文本分类模型学习的奠基指标，应不小于95%	正负例准确率＝标注正确的一级分类数据量/标注总数据量
	最小子类目准确率	验收关键指标，即标签类目的最细化分类判断准确率。最小子类目准确率是影响模型精度的强关注指标，通常要求90%～95%	最小子类目准确率＝标注正确的最小子类目数据量/标注总数据量
文本提取/摘要	关键词分级准确率	文本抽取类型标注中，对于关键词等级及对应等级的词汇选择标注，分级正确且词选正确，为最终正确，这也是关键词抽取类项目的最高准确要求，通常建议指标设定在80%～90%	关键词分级准确率＝对应等级词选正确数/所有等级对应词选数总和

表 13　文本数据标注/审核流程质量要素及测量方式（续）

适用项目类型	指标名称	指标定义及要求	计算逻辑
排序打分	排序分层一致率	文本排序打分类型的项目，有部分边界地带case难以精确评分，便通过排序分层来归类该case所属范畴。其操作方式为两个作业人员对一个case进行打分，如打分结果相邻，则判定排序分层一致。排序分层一致率设定在80%～95%，都属于合理范围	排序分层一致率＝同一case被两人标注结果相近的个数／同一case被两人标注的总数
通用指标	盲审一致率	对于主观性强的文本项目，通常采用盲审一致率来辅助准确率的评估，其操作方式为两个作业人员对一个case进行标注，如标注结果一致，则判定标注正确，一致率可以在一定程度上反映整体准确率。盲审一致率通常要求不低于70%	盲审一致率＝同一case被两人标注结果相同的个数/同一case被两人标注的总数

文本数据标注/审核测量方式说明见表 14。

表 14　文本数据标注/审核测量方式说明

测量方式		说明
抽样检查	定向抽检	随机抽检是面向抽检对象总任务池，设置一定条件（标注账号/任务日期/标注结果/标注轮次等）进行任务筛选，满足抽取条件的任务将被定向抽样检查
	随机抽检	随机抽检是抽检对象的总任务池中每个任务都有同等被抽中的可能，是一种完全依照机会均等的原则进行的抽样检查
多轮次	两轮	多轮次审核将一个标注任务互斥分发给2个作业人员，如2人判断结果一致，则默认该任务标注正确；如判断结果不一致，则任务自动流入质检池，由质检人员裁决其准确情况
评估	再抽样	评估是在抽检池中，设置一定条件（抽检日期/数量比例/数据来源等）进行任务筛选，满足抽取条件的任务将流入评估池进行再质检和评价估量

文本数据标注/审核质量检查规范见表 15。

表 15　文本数据标注/审核质量检查规范

资源	问题答案	检查人username	检查是否合格	质检人username	质检是否合格	验收人username	验收是否合格	验收建议
＜文本内容＞	{"item":{"src":" 三、劳动报酬\n 1 甲方愿意执行国家有关工资支付的规定，遵守最低工资保障制度。甲方工资发放日为每月____日，工资发放形式为直接发放／委托第三方代发。","templateID":"[redacted]type":"text-split","data":[{"type":"text","data":[{"type":"single-select","required":true,"description":"全文属性","options":["1","2","3"],"result							

7.2.2.4 标注/审核数据格式

文字数据处理支持 csv、url、doc 等常用交付格式，内含标注字段如 UID、垂类领域、关键词、标题、评论、前端原文链接等。

7.2.3 文本数据流通分发技术指标及测量

7.2.3.1 流通分发设定流程

参照 7.1.3.1。

7.2.3.2 流通分发工具要求

流通分发可采用多种方式和工具进行：

——API 接口：通过 API 接口进行数据传输；

——Url 访问：通过 url 链接对数据进行访问；

——FTP 传输：使用 ftp 传输工具进行数据传输；

——私有云传输：通过密钥访问私有云数据；

——其他工具：网盘等工具。

7.2.3.3 流通分发质量标准

流通分发质量由内容质量和传输质量构成。数据内容质量参照 7.2.1.3 采集质量标准和 7.2.2.3 标注/审核质量标准。传输质量参照相关传输方式技术，其中 API 接口标准参照 7.1.3.3 语音数据流通分发质量标准。

7.2.3.4 流通分发数据格式

——数据输入：csv、url、doc；

——数据输出：csv、url、doc。

7.3 图像数据服务技术

7.3.1 图像数据采集和预处理技术指标及测量

7.3.1.1 采集设定流程

参照 7.1.1.1。

7.3.1.2 采集工具要求

图像采集途径通常为人工采集。计算机视觉领域常有 2D、3D、点云、红外、双目深度等数据的采集类型。图像采集工具除常规手机外，不同需求下对硬件有特定要求，如计算机视觉领域的深度相机、红外相机、毫米波雷达、Xsens 手套组合、人体 3D 扫描仪等。

——普通设备：手机、普通相机、监控摄像头等，用于常规图片数据的采集。

——专业设备：专业相机、鱼眼相机等，用于特殊场景数据的采集。

——软件工具：图片采集软件，可以直接传输到客户的数据库，或上传数据平台，节省数据传输的时间。

7.3.1.3 采集质量标准

采集质量涉及采集服务成果质量和流程质量两方面。

图像数据采集流程质量要素及测量方式见表 16。

表 16 图像数据采集流程质量要素及测量方式

指标名称	指标定义及要求	计算逻辑
一次交付达成率	项目数据一次交付时准确率的达成情况,用于衡量项目的质量保证能力	一次交付达成率=(一次交付准确率/目标准确率)×100%
单次交付合格率	项目数据每次(按项目与业务约定的交付周期:日、周、月、数据包)交付时准确率的达成情况,用于评估交付能力	单次交付合格率=1-(项目单次交付不合格数量/项目单次交付数量)×100%
终审交付达成率	项目数据终审通过的数量,用于衡量项目的交付质量情况	终审交付达成率=(终审交付的合格数据量/目标交付合格的数据量)×100%
交付周期延时率	量级要求项目:实际交付周期与目标交付周期的时间差。用于评估交付能力	交付延时率=[(实际交付周期-约定交付周期)/约定交付周期]×100%

图片采集项目类型会按照不同的应用场景有不同的用途,例如应用于人脸识别的人脸图像采集,新零售场景的商品采集、人体姿态采集,应用于自动驾驶场景的道路行人图片采集等等。每一个子类型的采集质检质量侧重点不同,如无特殊要求,可参照正确性检验规范来要求。正确性检验规范应包含数据的场景要求和图像信息要求两个维度,场景要求主要包含是场景内包含要求的信息准确性、数量要求、位置要求、时效等,图像信息包含图像分辨率、清晰度、图片大小等。

a) 待标注数据——不带预标注结果:

——csv 第 1 列存放图片地址,列名不能为空,可自定义(中英文均可);

——第 2 列列名不能为空,可自定义(中英文均可)。

资源	问题答案
<图像链接>,例如<https://XXXXXXX.jpg>	

b) 待标注数据——带预标注结果:

——csv 第 1 列存放图片地址,列名不能为空,可自定义(中英文均可);

——第 2 列存放 json 结构的预标注结果,列名不能为空,可自定义(中英文均可);

——json 结构的预标注结果:可先从队列里标注一个题目,获取该题目的标注结果,将预标注的结果替换为对应 json 结构,再上传到队列,则每个内容自带预标注结果。

资源	问题答案
<图像链接>,例如<https://XXXXXXX.jpg>	{"item":{"src":"// 8[redacted]13"},"templateID":"68[redacted]type":"graph- mark","data":[{"regionType":"image","data":[{"type":"single- select","required":true,"description":"天气","options":["晴天","阴天","雨天"],"result":"晴天 "},{"type":"single-select","required":true,"description":"时间","options":["白天","夜间 "],"result":"白天"}]},{"regionType":"rect","regionID":1,"region":{"tl":{"x":0,"y":- 1},"br":{"x":250,"y":225}},"data":[{"type":"single-select","required":true,"description":"交通工具 类型","options":["轿车","越野车","大卡车","公交车"],"result":"轿车"},{"type":"single-

7.3.1.4 采集数据格式

常见的图像格式有 JPEG、TIFF、RAW、BMP、GIF、PNG 等。

7.3.2 图像数据标注/内容审核技术指标及测量

7.3.2.1 标注/审核设定流程

参照 7.1.2.1。

7.3.2.2 标注/审核工具要求

图像数据标注/审核流程功能要求见表 17。

表 17 图像数据标注/审核流程功能要求

图片处理类型	模板元素	功能要求	适用项目类型
通用类	图片缩放/旋转/移动、亮度/对比度/饱和度调整等	图片类数据处理基础功能要求，对图片进行基础缩放、旋转等操作，包括但不限于增删标签并调整其颜色和显隐，一键复原或查看原图，撤销及自动保存等功能	通用
OCR检测/识别类	矩形/多边形框	对需要标注的目标实体用矩形框或多边形框进行框选，以此和周围进行边界区分，要求需同时实现增、删、调整框，切换框的填充色及透明度等	左右手框检测标注
	类别标签	对已用矩形或多边形框标注的目标增删类别标签，以区分图片中不同目标属性	
关键点类	连线设置	对于标注过程中的一组关键点进行自动连线，并在不同区间段进行闭环隔离以区分独立关键点组，如1-8-1，9-16-9等	人脸轮廓/瞳孔虹膜/精细关键点标注
	椭圆拟合	适用于椭圆形或类椭圆形区域进行关键点定位时，自动由4点拟合椭圆圆弧，代替默认连线折线段	
	点序号/显隐调整	可对连续性关键点可自定义序号起始及显隐，以判断局部特殊点位是否精准	
精细化抠图/分割类	正向画笔、负向画笔（擦除）	正向画笔针对画布上的目标拖动进行绘制线条、勾勒轮廓以便后续填充区域，产生与原图叠加的分割效果图；负向画笔即为擦除笔，针对已绘制或已分割区域进行边缘修正、区域擦除等动作	人脸局部分割/人体服饰分割/车辆车窗背景分割
	涂色工具	将已绘制轮廓进行区域涂色，分前景颜色、背景颜色等，实现原图不同区域着色分割效果	
	移动/拖拽按钮	对已用画笔进行分割的部分进行移动，对比原图判断分割效果	

标注工具及作用：

——数据配置：支持一题一图，或一题多图。一题多图等于图片连续帧；

——允许图片旋转/翻转：开启后，标注图片可进行上下左右旋转及镜像翻转，旋转及标注后的结果，会继续流入到后续环节；

——按选项区分颜色：可自定义每个选项的 rgb 色值；

——显示打点顺序：按“R”键，可展示框/多边形/线的落点顺序；

——TrackId：可对多个标注结果备注一个“ID 编号”，便于识别，如多边形标记汽车，点功能标记轮胎，trackid 都标记为 1，代表一组标注结果；

——禁止修改预标注结果：开启后不可修改或删除预标注结果；

——最小框限制：开启后，新建框、调整框无法小于该尺寸；
——禁止标到图片外：开启后不可标注到图像外；
——点组功能：自定义几个点为一组的方式进行标注。

7.3.2.3 标注/审核质量标注

图像数据标注/审核质量要素及测量方式见表 18。

表 18 图像数据标注/审核质量要素及测量方式

适用类型	指标名称	指标定义及要求	计算逻辑
OCR检测及识别、关键点标注等	最小单位准确率	验收关键指标，即图片标注的最小颗粒单位准确率（如框、点、折线、帧数等），最小单位准确率是需求方强关注交付指标，具体指标值要求根据业务场景确定	最小单位准确率=（标注正确的最小单位正确数/最小单位标注总数）×100%
图片筛选/分类/清洗等	标签正负例准确率	多用于二分类场景，属于指定标签下的数据即为正例，不属于则为负例，标签正负例准确率是筛选/清洗等分类场景下定义数据质量的关键指标，具体指标值要求根据业务场景确定	标签正负例准确率=（标注正确的正例数量/标注总数）×100%
属性判断类（含较强主观因素）	多轮次一致率	常用于视频抽帧数据属性判断场景，关注多轮标注模式下不同轮次标注一致率，验证该类数据主观程度及规则拉齐难易度，具体指标值要求根据业务场景确定	多轮次一致率=（n轮标注后一致数量/单轮总数）×100%

图像数据标注/审核测量方式说明见表 19。

表 19 图像数据标注/审核测量方式说明

测量方式		说明
抽样检查	定向抽检	随机抽检是面向抽检对象总任务池，设置一定条件（标注账号/任务日期/标注结果/标注轮次等）进行任务筛选，满足抽取条件的任务将被定向抽样检查
	随机抽检	随机抽检是抽检对象的总任务池中每个任务都有同等被抽中的可能，是一种完全依照机会均等的原则进行的抽样检查
多轮次	两轮	多轮次审核将一个标注任务互斥分发给2个标注人员，如2人判断结果一致，则默认该任务标注正确；如判断结果不一致，则任务自动流转到第3轮，由第3人（常为质检人员）最终裁决其准确情况
埋点标注	埋点质检	一批数据中包含一部分原始带有标注结果的数据，并均匀分布到数据样本中，通过单轮标注后验证埋点数据的标注质量，从而验证整体的数据质量

7.3.2.4 标注/审核数据格式

图片数据处理支持 csv、url、psd 等常用交付格式，内含标注字段如关键点（组）坐标、矩形框四点坐标及其标签类别等。

7.3.3 图像数据流通分发技术指标及测量

7.3.3.1 流通分发设定流程

参照 7.1.3.1。

7.3.3.2 流通分发工具要求

流通分发可采用多种方式和工具进行：

——API 接口：通过 API 接口进行数据传输；

——Url 访问：通过 url 链接对数据进行访问；

——FTP 传输：使用 ftp 传输工具进行数据传输；

——私有云传输：通过密钥访问私有云数据；

——其他工具：网盘等工具。

7.3.3.3 流通分发质量标准

流通分发质量由内容质量和传输质量构成。数据内容质量参照 7.3.1.3 采集质量标准和 7.3.2.3 标注/审核质量标准。传输质量参照相关传输方式技术规范，其中 API 接口标准参照 7.1.3.3 语音数据流通分发质量标准。

7.3.3.4 流通分发数据格式

流通分发数据格式参照 7.3.1.4 采集数据格式和 7.3.2.4 标注/审核数据格式。

7.4 视频数据服务技术

7.4.1 视频数据采集技术指标及测量

7.4.1.1 采集设定流程

参照 7.1.1.1。

7.4.1.2 采集工具要求

图像采集工具除常规手机外，不同需求下对硬件有特定要求，如计算机视觉领域的深度相机、红外相机、毫米波雷达、Xsens 手套组合、人体 3D 扫描仪等。

——普通设备：手机、普通摄相机、监控摄像头等，用于常规视频数据的采集。

——专业设备：专业摄像设备，用于特殊场景视频数据的采集。

——网络采集：从网上收集整理视频数据。

7.4.1.3 采集质量标准

参照 7.3.1.3。

7.4.1.4 采集数据格式

常见的视频格式有 MPEG、MP4、AVI、3GP、RM(RMVB)、WMV、FLV(F4V)等。

7.4.2 视频数据标注/内容审核技术指标及测量

7.4.2.1 标注/审核设定流程

参照 7.1.2.1。

7.4.2.2 标注/审核工具要求

视频数据标注/审核工具平台功能要求见表 20。

表 20 视频数据标注/审核工具平台功能要求

视频处理类型	模板元素	功能要求	适用项目类型
视频选择类	单选标签/复选框/下拉框	视频类数据处理基础需求，不对本身数据进行修改，仅做选择类的操作;包括打标签、打分等	判断、分类、清洗、评测、打分
智能剪辑/切分	视频剪辑工具	按指定要求在时间轴上进行切分，要求工具可实现开始时间、结束时间、或者按照要求的时间打点等	视频剪辑/切分
文本转写	矩形/多边形框/类别标签	通过矩形框/多边形框框住视频中的文本，并自动将框选的文本进行转录，确保文本正确性	视频文本转写
内容解析	文本输入框	根据内容进行分析，给予分析报告或分析说明，通过文本输入框录入文字信息	视频解析
画面遮挡	马赛克	根据规则对视频中的特定画面进行遮挡	视频打码

7.4.2.3 标注/审核质量标准

视频数据标注/审核质量要素及测量方式见表 21。

表 21 视频数据标注/审核质量要素及测量方式

适用项目类型	指标名称	指标定义及要求	计算逻辑
视频判断、分类、清洗、评测、打分	标签正负例准确率	多用于视频打标签场景，属于指定标签下的数据即为正例，不属于则为负例，标签正负例准确率是筛选/清洗等分类场景下定义数据质量的关键指标，具体指标值要求根据业务场景确定	标签正负例准确率=(标注正确的正例数量/标注总数)×100%
内容解析	评估准确率	适用于视频解析项目，按照标准中给到的内容要素，评估内容解析的内容是否符合需求方预期，具体指标值要求根据业务场景确定	评估准确率=(满足条件的内容数量/标注总数)×100%
文本转写 智能剪辑/切分 画面遮挡	最小单位准确率	验收关键指标，即视频标注的最小颗粒单位准确率(如框、点、视频抽帧、转写文字等)，最小单位准确率是需求方强关注交付指标，具体指标值要求根据业务场景确定	最小单位准确率=(标注正确的最小单位正确数/最小单位标注总数)×100%
视频打分、分类	多轮次一致率	适用于主观判断类视频抽帧数据，多轮标注模式下不同轮次标注一致率，通过用于验证该类数据主观程度及规则拉齐难易度，具体指标值要求根据业务场景确定	多轮次一致率=(n轮标注后一致数量/单轮总数)×100%

表 22　视频数据标注/审核质量测量方式说明

测量方式		说明
抽样检查	定向抽检	随机抽检是面向抽检对象总任务池，设置一定条件（标注账号/任务日期/标注结果/标注轮次等）进行任务筛选，满足抽取条件的任务将被定向抽样检查
	随机抽检	随机抽检是抽检对象的总任务池中每个任务都有同等被抽中的可能，是一种完全依照机会均等的原则进行的抽样检查
多轮次	两轮	多轮次审核将一个标注任务互斥分发给2个作业人员，如2人判断结果一致，则默认该任务标注正确；如判断结果不一致，则任务自动流入质检池，由质检人员裁决其准确情况
评估	再抽样	评估是在抽检池中，设置一定条件（抽检日期/数量比例/数据来源等）进行任务筛选，满足抽取条件的任务将流入评估池进行再质检和评价估量

7.4.2.4　标注/审核数据格式

组合模版对需要展示的数据进行配置，通过变量名（src 字段）与上传 CSV 的表头名称进行匹配展示数据。

页面模版与数据的关系图如图 7 所示。

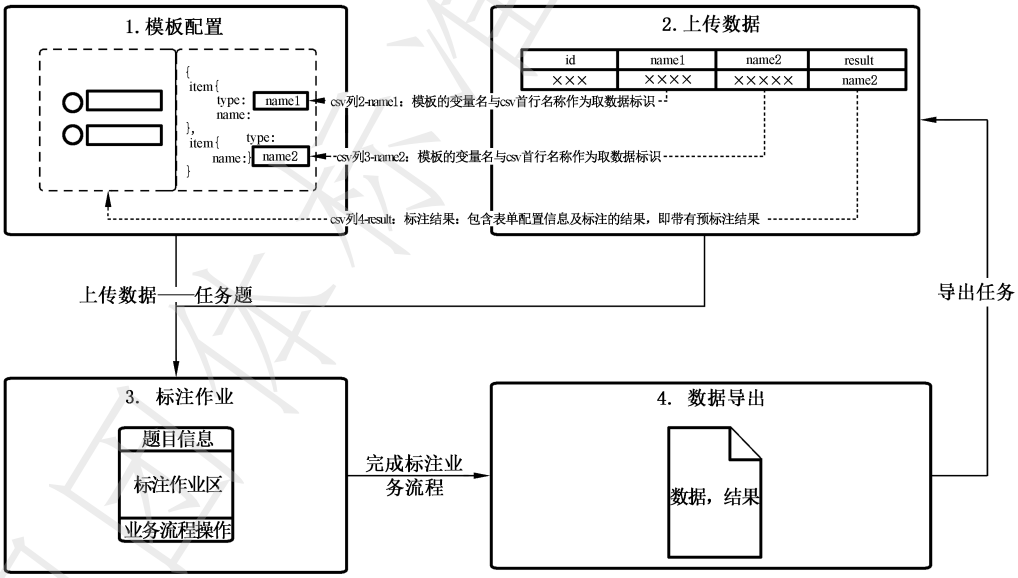

图 7　页面模版与数据的关系图

组合数据的展示方式配置：这里以视频、标题、标题内容这三种待标注数据为例：

格式	字段
数据输入：csv、url、mp4、mov、wmv、avi等	视频ID、视频
数据输出：csv、url、mp4、mov、wmv、avi等	标签类别、markTime位置坐标、endTime位置坐标、截取片段时长、截取片段占总视频比例、打分等

7.4.3 视频数据流通分发技术指标及测量

7.4.3.1 流通分发设定流程

参照 7.1.3.1。

7.4.3.2 流通分发工具要求

可采用多种方式和工具进行：

——Url 访问：通过 url 链接对数据进行访问；

——API 传输：通过 API 接口进行数据传输；

——FTP 传输：使用 ftp 传输工具进行数据传输；

——私有云传输：通过密钥访问私有云数据；

——其他工具：网盘等工具。

7.4.3.3 流通分发质量标准

流通分发质量由内容质量和传输质量构成。数据内容质量参照 7.4.1.3 采集质量标准和 7.4.2.3 标注/审核质量标准。传输质量参照相关传输方式技术规范，其中 API 接口标准参照 7.1.3.3 语音数据流通分发质量标准。

7.4.3.4 流通分发数据格式

流通分发数据格式参照 7.4.1.4 采集数据格式和 7.4.2.4 标注/审核数据格式。

7.5 点云数据服务技术

7.5.1 点云数据采集和预处理技术指标及测量

7.5.1.1 采集设定流程

参照 7.1.1.1。

7.5.1.2 采集工具要求

采集工具要求如下。

——三维激光雷达扫描：激光照射到物体表面时，所反射的激光会携带方位、距离等信息。根据这种方法得到的点云，一般具有 x、y、z 坐标值以及激光反射强度（Intensity）4 种信息。

——照相机扫描：这种方法得到的点云，除 x、y、z 坐标以为还具有颜色信息。

——逆向工程：在三维模型的表面进行采点，得到相应的点云。具有点所在平面的法向量信息。

7.5.1.3 采集质量标准

采集质量涉及采集服务成果质量和流程质量两方面。

点云数据采集流程质量要素及测量方式见表 23。

表 23　点云数据采集流程质量要素及测量方式

指标名称	指标定义及要求	计算逻辑
一次交付达成率	项目数据一次交付时准确率的达成情况,用于衡量项目的质量保证能力	一次交付达成率=(一次交付准确率/目标准确率)×100%
单次交付合格率	项目数据每次(按项目与业务约定的交付周期:日、周、月、数据包)交付时准确率的达成情况,用于评估交付能力	单次交付合格率=1−(项目单次交付不合格数量/项目单次交付数量)×100%
终审交付达成率	项目数据终审通过的数量,用于衡量项目的交付质量情况	终审交付达成率=(终审交付的合格数据量/目标交付合格的数据量)×100%
交付周期延时率	量级要求项目:实际交付周期与目标交付周期的时间差。用于评估交付能力	交付延时率=((实际交付周期−约定交付周期)/约定交付周期)×100%

点云采集项目类型目前应用于自动驾驶领域比较多,采集初期需要根据算法侧需求用特定的激光雷达、摄像头、传感器等进行采集。每一个子类型的采集质检质量侧重点不同,如无特殊要求,可参照正确性检验规范来要求。正确性检验规范应包含数据的场景要求和图像信息要求两个维度,场景要求主要包含是场景内包含要求的信息准确性、数量要求、位置要求、时效等,图像信息包含点云成像效果、清晰度等。

7.5.1.4　采集数据格式

常见的点云数据格式有:PTS、XYZ、LAS、PLY。

普通点云数据和开启 2D&3D 融合后的数据结构是不一样的。

普通点云数据:

a)　无预标注结果:csv 第 1 列:表头固定为“point_cloud”,存放点云数据链接,仅支持 .pcd 格式文件。无预标注结果 CSV 示例:

point_cloud	
<点云链接>	

b)　有预标注结果:

——csv 第 1 列:表头固定为“point_cloud”,存放点云数据链接,仅支持 .pcd 格式文件;

——csv 第 2 列:表头固定为“pre_result”,存放预标注结果,预标注结果为 json 结构。

有预标注结果 CSV 示例:

point_cloud	pre_result
<点云链接>	

7.5.2　点云数据标注/内容审核技术指标及测量

7.5.2.1　标注/审核设定流程

参照 7.1.2.1。

7.5.2.2　标注/审核工具要求

点云数据标注/审核工具平台功能要求见表 24。

表 24　点云数据标注/审核工具平台功能要求

处理类型	模板元素	功能要求	适用项目类型
通用类	点云图像缩放/旋转/移动、亮度/对比度/饱和度调整等	点云数据处理基础功能要求，对图片进行基础缩放、旋转等操作，包括但不限于增删标签并调整其颜色和显隐，一键复原或查看原图，撤销及自动保存等功能	通用
点云标注	矩形/多边形框	对需要标注的目标实体用矩形框或多边形框进行框选，以此和周围进行边界区分，要求需同时实现增、删、调整框，切换框的填充色及透明度等	纯点云标注、联合标注、融合标注
	类别标签	对已用矩形或多边形框标注的目标增删类别标签，以区分图片中不同目标属性	
	椭圆拟合	适用于椭圆形或类椭圆形区域进行关键点定位时，自动由 4 点拟合椭圆圆弧，代替默认连线折线段	
精细化抠图/分割类	正向画笔、负向画笔（擦除）	正向画笔针对画布上的目标拖动进行绘制线条、勾勒轮廓以便后续填充区域，产生与原图叠加的分割效果图；负向画笔即为擦除笔，针对已绘制或已分割区域进行边缘修正、区域擦除等动作	点云语义分割
	涂色工具	将已绘制轮廓进行区域涂色，分前景颜色、背景颜色等，实现原图不同区域着色分割效果	
	移动/拖拽按钮	对已用画笔进行分割的部分进行移动，对比原图判断分割效果	

点云分类和与其他功能同时开启。

a）　3D 框模式：

——一题多帧：等于点云连续帧功能；

——是否允许修改 TrackId：点云类标注默认开启 TrackID，可系统自动分配 id 编号，或自己修改 id 编号；

——开启 2D3D 映射：即 2&3D 融合标注；

——3D 映射到 2D 为 3D 框：点云上标注 3D 框后，2D 图上会映射展示 3D 框；

——3D 映射到 2D 为 2D 框：点云上标注 3D 框后，2D 图上会映射展示 2D 框；

——允许修改 2D 结果：点云上标注 3D 框后，可对映射在 2D 图上的 2D 框进行形状和位置的修改；

——类型属性：每个题目及题目下的标签，支持作为公共属性或 3D/2D 独有属性；

b）　分割模式：

——多问题选项互斥：在多个问题选项内单选；

——2D 参考图：开启后作业界面可看到 2D 参考图，导入数据时需确保导入图片链接和 calib 文件链接。

c）　线模式：

——2D 参考图：开启后作业界面可看到 2D 参考图，导入数据时需确保导入图片链接和 calib 文件链接；

——是否允许修改 TrackId：点云类标注默认开启 TrackID，可系统自动分配 id 编号，或自己修改 id 编号。

7.5.2.3 标注/审核质量标准

自动驾驶中的常见标注类型为点云 3D 框标注、3D 点云语义分割、2D3D 融合标注、点云车道线等。点云数据标注/审核质量要素及测量方式见表 25。

表 25 点云数据标注/审核质量要素及测量方式

常见类型	质量指标	定义	指标内涵
点云 3D 框	框准确率	框准确率=(合格框数/总框数)×100%	——完整:将目标物体完整标入框内,3D 标注框大小合理; ——角度:点云 3D 标注框无明显的角度偏移; ——预测:合理判断出点云不全的标注对象的真实尺寸大小; ——类别:标注对象分类无错误,相关属性无错误; ——一致性:连续帧中要求同物体 ID 和尺寸一致性
点云车道线	线准确率	线准确率=(合格线数/总线数)×100%	——贴合:点云车道线贴合实际场景; ——类别:标注对象分类无错误,相关属性无错误
3D 点云语义分割	帧准确率,框准确率,点准确率	帧准确率=(合格帧数/总帧数)×100%; 框选准确率=(合格框数/总框数)×100%; 点准确率=(合格点属性/总点数)×100%	——无漏标:无漏标点集、标签; ——目标边界准确:不同目标的边界区分准确; ——调整视角:检查过程中多视角调整切换,找到合适视角确认不同类型目标边界
2D3D 融合标注	2D 框框准确率,3D 框框准确率,帧准确率	2D 框框准确率=(合格框数/总框数)×100%; 3D 框框准确率=(合格框数/总框数)×100%; 帧准确率=(合格帧数/总帧数)×100%	——框准确率参考点云 3D 框内涵; ——帧准确率=(合格帧数/总帧数)×100%

7.5.2.4 标注/审核数据格式

7.5.2.4.1 PCD 格式简介

PCD 全称 Point Cloud Data,是一种存储点云数据的常见文件格式。PCD 存储格式是 PCL 库官方指定格式,典型的为点云量身定制的格式。文件格式有文本和二进制两种格式。PCD 格式具有文件头,用于描述点云的整体信息:点云尺寸、维数和数据类型。数据本体部分由点的坐标和其他维度构成。文本模式下以空格做分隔符。

7.5.2.4.2 其他点云存储格式

点云数据存储格式说明见表 26。

表 26　点云数据存储格式说明

格式	说明
PTS	点云文件格式是最简便的点云格式，直接按 X、Y、Z 顺序存储点云数据，可以是整型或者浮点型
XYZ	一种文本格式，存储点的三维坐标和法向量，数字间以空格分隔
LAS	是激光雷达数据（LiDAR），提供一种开放的格式标准，允许不同的硬件和软件提供商输出可互操作的统一格式
PLY	表示多边形的文件格式

7.5.2.4.3　点云数据文件头格式

点云数据文件头格式说明见表 27。

表 27　点云数据文件头格式说明

属性	说明
VERSION	PCD 文件的版本号
FIELDS	每个点所包含的维度或字段
SIZE	每个数据维度所占据的字节
TYPE	每个数据维度的数据类型（I：有符号整型、U：无符号整型、F：浮点型）
COUNT	每个数据维度上包含的元素个数，默认为 1
WIDTH	点云数据集的宽度 ——有序点云：一行中点的数目； ——无序点云：总点数同 POINTS
HEIGHT	点云数据集的高度： ——有序点云：一列中点的数目； ——无序点云：1
VIEWPOINT	采集点的视角，平移 + 四元数，默认是 0 0 0 1 0 0 0
POINTS	点的数量
DATA	数据存储格式（ascii | binary | binary_compressed ）

7.5.2.4.4　点云相机投影参考

相机定义了场景中的可视区域，相机有自己的位置（position）、朝向（lookAt）和视锥体（可视域）。创建的图形一定要放在相机的视锥体中才能看得见。视锥体与相机的类型、相机的位置和朝向都有关。

相机主要有两种：正投影相机和透视投影相机。正投影的投影长度与几何体倾斜角度有关；透视投影的投影长度与几何体位置有关，如图 8 所示。

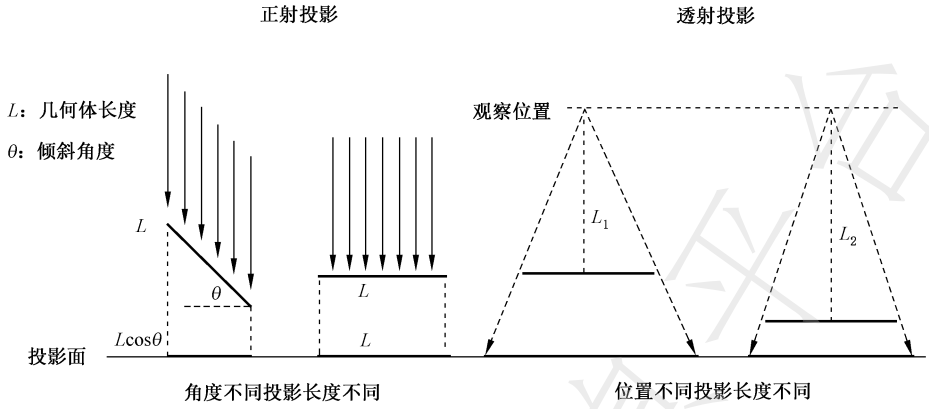

图 8　点云相机投影示例

7.5.3　点云数据流通分发技术指标及测量

7.5.3.1　流通分发设定流程

参照 7.1.3.1。

7.5.3.2　流通分发工具要求

流通分发可采用多种方式和工具进行：

——API 接口：通过 API 接口进行数据传输；

——Url 访问：通过 url 链接对数据进行访问；

——FTP 传输：使用 ftp 传输工具进行数据传输；

——私有云传输：通过密钥访问私有云数据；

——其他工具：网盘等工具。

7.5.3.3　流通分发质量标准

流通分发质量由内容质量和传输质量构成。数据内容质量参照 7.5.1.3 采集质量标准和 7.5.2.3 标注/审核质量标准。传输质量参照相关传输方式技术规范，其中 API 接口标准参照 7.1.3.3 语音数据流通分发质量标准。

7.5.3.4　流通分发数据格式

流通分发数据格式参照 7.5.1.4 采集数据格式和 7.5.2.4 标注/审核数据格式。

参 考 文 献

[1] GB/T 7408 数据元和交换格式信息交换日期和时间表示法
[2] GB/T 18391 信息技术 元数据注册系统(MDR)
[3] GB/T 33767(所有部分) 信息技术 生物特征样本质量
[4] GB/T 34077.1 基于云计算的电子政务公共平台管理规范 第1部分:服务质量评估
[5] GB/T 37988 信息安全技术 数据安全能力成熟度模型
[6] GB/T 38667 信息技术 大数据 数据分类指南
[7] TC260-PG-20212A 网络安全标准实践指南 网络数据分类分级指引