ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T XXXXX—XXXX

高质量数据集 格式要求

High-quality dataset—Format requirements

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

国家市场监督管理总局
国家标准化管理委员会　发布

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。

本文件由全国数据标准化技术委员会（SAC/TC 609）提出并归口。

本文件起草单位：中国电子技术标准化研究院华东分院、中国电子信息产业发展研究院、中国移动通信集团有限公司、中国电子技术标准化研究院、中国科学院计算技术研究所、国家数据发展研究院、中电数据产业集团有限公司、国务院国有资产监督管理委员会研究中心、交通运输部公路科学研究所、北京大学、公安部第三研究所、中国石油天然气集团有限公司、中国石油化工集团有限公司、中国交通建设集团有限公司、国家能源投资集团有限责任公司信息技术分公司、国家电网有限公司大数据中心、中国南方电网有限责任公司、国家石油天然气管网集团有限公司、浦江国家实验室、工业和信息化部电子第五研究所、中国联合网络通信集团有限公司、中国电信集团有限公司、中国质量认证中心有限公司、煤炭科学研究总院有限公司、中国稀土集团有限公司、华为技术有限公司、科大讯飞股份有限公司、阿里巴巴（中国）有限公司、北京智源人工智能研究院、北京百度网讯科技有限公司、深圳市腾讯计算机系统有限公司、商业信用中心、中国信息通信研究院、北京智网数科技术有限公司、石化盈科信息技术有限责任公司、中国交通信息科技集团有限公司、中移动信息技术有限公司、中移互联网有限公司、国家电投集团数字科技有限公司、中石油（北京）数智研究院有限公司、联通数据智能有限公司、上海库帕思科技有限公司、上海信投智能科技股份有限公司、航天科工网络信息发展有限公司、中国邮政储蓄银行股份有限公司、中电金信软件有限公司、江苏省大数据管理中心、内蒙古自治区大数据中心、江西省大数据中心、四川省卫生健康信息中心（四川省健康医疗大数据中心）、北京大学（天津滨海）新一代信息技术研究院、国家开放大学、杭州数美科技有限公司、福建省大数据集团有限公司、湖北大数据集团数据开发有限公司、南京南瑞继保工程技术有限公司、南京南瑞瑞中数据股份有限公司、中通服网盈科技有限公司、北京海天瑞声科技股份有限公司、广州数字健康科技有限公司、安徽飞数信息科技有限公司、卡奥斯工业智能研究院（青岛）有限公司、杭州市临安区大数据管理服务中心、国网河南省电力公司经济技术研究院、软通智慧科技有限公司、烽火通信科技股份有限公司、太极计算机股份有限公司、复旦大学、同方知网数字科技有限公司、中移雄安信息通信科技有限公司、数据堂（北京）科技股份有限公司、《智慧中国》杂志社有限责任公司、河南金盾信安检测评估中心有限公司、河南省泛物网络科技有限公司、信阳市璀璨科技有限责任公司、睿尔曼智能科技（北京）有限公司、北京银河通用机器人股份有限公司、中兴通讯股份有限公司、浪潮电子信息产业股份有限公司、国网山东省电力公司、蔚来汽车科技（安徽）有限公司、贵州大数据产业集团有限公司、杭州市数据集团有限公司、厦门赛西科技发展有限责任公司、云基华海信息技术股份有限公司、国网江苏省电力有限公司、国网江苏省电力有限公司、广东省人民医院、辽宁省电子信息产品监督检验院、数字宁波科技有限公司、杭州景联文科技有限公司、北京星河智源科技有限公司、山西集智数据服务有限公司、山东未来集团有限公司、广州维视达数字科技有限公司、厦门身份宝网络科技有限公司、上海森栩医学科技有限公司、北京中数睿智科技有限公司、江苏中堃数据技术有限公司。

本文件主要起草人：王为中、韩冰、张欢、吴坤、郭嘉丰、赵鹏飞、张群、廖华明、王超、温晓君、苏越阳、李天舒、时晓光、郭祎萍、连森、黄吉海、刘怡林、周艳芳、王亚沙、赵俊峰、马连韬、付文豪、谭瑾、李成博、蒋楠、刘陈宇、刘学忠、张黎明、周春雷、赵翔宇、贾蕾、李珍翔、程广明、李金夏、施晓辉、王锋、程健、骆意、张建中、武光城、谢卫军、罗腾、赵丽丽、王鑫、刘俊华、吴峥、李世奇、刘颖、刘广、杨二龙、邱泳钦、刘煜宏、肖邱勇、王雅琴、李荪、曹峰、姜辉、朱江涛、索寒生、

严龙云、王晶、李亚楠、梁小涛、杨山、王海波、薛健、刘速、潘登、谭晓坤、胡力旗、邓成龙、汪睿棋、王兴旺、林云峰、张冲、袁芮、李娜、刘超、代勤、袁小乐、沈明辉、向海平、周志华、邱会丽、孔亚文、蔡斯博、李静、孙晖、丁斌、张毅、周季峰、张锦辉、李凡、李科、黄宇恒、葛海龙、王培养、王宇、申中一、戴斌、王圆圆、林镇阳、陈刚、李佳忆、何震瀛、张庆国、杨彭年、齐红威、张挺、梁宏、何少英、马盼、郑随兵、王鹤、吴德亮、陈曦、邵志敏、范瑞、杨坤燚、张凯、邱旭东、鲁胜强、夏飞、王鹏飞、杨小红、王俊吉、李晓儿、刘云涛、严长春、庞俊奇、徐小传、彭荣、陈颖、温冬梅、韩涵、魏清。

引　言

当前，随着新一代信息技术持续快速发展，人工智能正加速融入各行业领域，赋能实体经济高质量发展。高质量数据集是开发和训练人工智能模型的重要支撑，能够提高模型精度与可解释性、减少训练时长，已经成为人工智能发展的核心要素。目前，在我国高质量数据集建设推进过程中，存在数据集格式不规范、不统一的问题。该问题不利于通过统一接口（或程序）对数据集进行读取、使用，进而阻碍数据集流通、应用。制定高质量数据集格式要求，明确其元数据及表示方法，规定数据标识、内容、标注、版本、授权、来源等方面要求，对于促进高质量数据集流通、应用，有力支持人工智能模型开发和训练，更好赋能经济社会发展至关重要。

高质量数据集 格式要求

1 范围

本文件规定了高质量数据集的元数据及其表示方法。

本文件适用于指导建设、管理和加工高质量数据集。

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 7408.1-2023 日期和时间 信息交换表示法 第1部分：基本原则

NDI-TR-2025-04 数据基础设施 标识管理规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

高质量数据集 high-quality dataset

经过采集、加工等数据处理，可直接用于开发和训练人工智能模型，能有效提升模型性能的数据的集合。

[来源：20255407-T-907，3.3.27]

3.2

元数据 metadata

定义和描述特定数据的数据，提供了关于数据的结构、特征和关系的信息，有助于组织、查找、理解、管理数据。

[来源：20255407-T-907，3.3.17]

3.3

值域 value domain

允许值的集合。

[来源：GB/T 18391.1-2009，3.3.38]

3.4

数据标注 data labeling；data annotation

给数据样本指定目标变量和赋值的过程。

[来源：20255407-T-907，3.4.13]

4 缩略语

下列缩略语适用于本文件。

URL：统一资源定位符（Uniform Resource Locator）
ISBN：国际标准书号（International Standard Book Number）
UTF-8：8 位统一码转换格式（Unicode Transformation Format - 8-bit）

5 元数据属性

每个元数据用 7 个属性描述，属性名及其定义见表 1。其中，数据类型包括日期型、布尔值、字符串、数组、对象、空值等；对于值域，自由文本指用户可以根据需要输入自定义文本内容；对于数据填充约束，“M”表示必须填写，“O”表示可选填写。

表1 元数据属性

序号	属性名	定义
1	中文名称	元数据的中文名称
2	英文名称	元数据的英文名称
3	定义	元数据含义的解释
4	数据类型	元数据的有效值类型
5	值域	元数据所允许值的集合
6	数据填充约束	元数据的数据填写约束
7	备注	元数据的附加注释

6 数据集元数据

本文件分数据标识、关联数据标识、数据内容、标注信息、原始时间、最后修改时间、数据版本、授权类型、来源类型、来源详情、生成数据标志给出数据集元数据，如图 1 所示。其中，数据内容元数据分模态类型、内容给出；标注信息元数据分标签、标注方式、标注工具、标注人员类型给出。本文件以 JSON 格式给出了高质量数据集的元数据示例，详见附录 A。

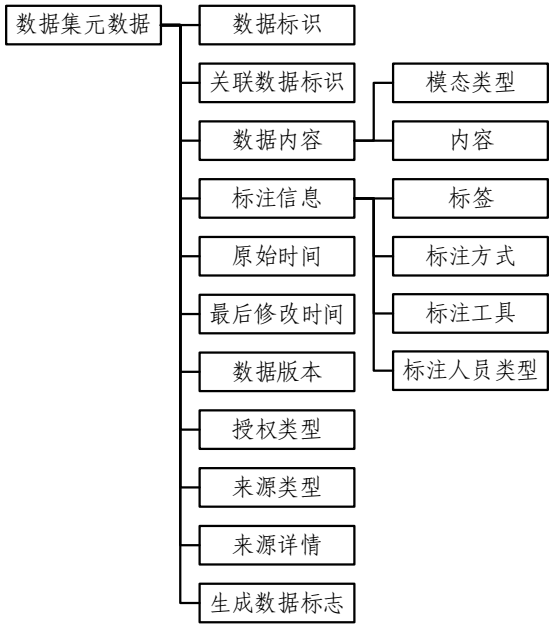

图1 数据集元数据示意图

6.1 数据标识

中文名称：数据标识
英文名称：id
定义：数据的全域唯一标识
数据类型：字符串
值域：应符合 NDI-TR-2025-04 的要求
数据填充约束：M
备注：无。

6.2 关联数据标识

中文名称：关联数据标识
英文名称：rid
定义：与当前数据存在明确关联关系的其他数据的数据标识的集合
数据类型：数组
值域：应符合 NDI-TR-2025-04 的要求
数据填充约束：O
备注：关联数据标识可以是一个或多个数据标识的有序列表。

6.3 数据内容

中文名称：数据内容
英文名称：data_content
定义：数据的具体内容
数据类型：数组
值域：应符合第 7 章的要求
数据填充约束：M
备注：数据内容可以是单条数据、同一模态的多条数据组合或不同模态的多条数据组合。

6.4 标注信息

中文名称：标注信息
英文名称：annotation
定义：数据的标注信息
数据类型：对象
值域：应符合第 8 章的要求
数据填充约束：O
备注：无。

6.5 原始时间

中文名称：原始时间
英文名称：original_time
定义：数据最初出现或创建的时间
数据类型：日期型
值域：应符合 GB/T 7408.1-2023 的要求

数据填充约束：M
备注：表示数据的原始创建或首次出现时间,而非采集时间。

6.6 最后修改时间

中文名称：最后修改时间
英文名称：last_modified_time
定义：数据最后一次被修改或加工的时间
数据类型：日期型
值域：应符合 GB/T 7408.1-2023 的要求
数据填充约束：M
备注：如果数据未经修改或加工,则与原始时间相同。

6.7 数据版本

中文名称：数据版本
英文名称：version
定义：数据的版本号
数据类型：字符串
值域：应符合《语义化版本（Semantic Versioning）》的要求
数据填充约束：M
备注：无。

6.8 授权类型

中文名称：授权类型
英文名称：license
定义：数据的授权类型
数据类型：字符串
值域：开源、公共授权、商业授权、仅内部、其他
数据填充约束：M
备注：无。

6.9 来源类型

中文名称：来源类型
英文名称：source
定义：数据的来源类型
数据类型：字符串
值域：互联网、图书、论文、报告、标准、专利、政府文件、组织机构等
数据填充约束：M
备注：此处列举的是数据的常见来源类型。当数据的来源类型不在列举范围内时，可采用符合实际的其他来源类型表示。

6.10 来源详情

中文名称：来源详情
英文名称：source_details

定义：来源类型的详细说明
数据类型：字符串
值域：自由文本
数据填充约束：M
备注：如互联网数据的 URL、图书的 ISBN 号和页码、论文的发表信息、报告的发布信息、标准的标准号、专利的专利号、政府文件的发文字号、组织机构的具体名称等。

6.11 生成数据标志

中文名称：生成数据标志
英文名称：generated_data_indicator
定义：数据是否为生成数据的标志
数据类型：布尔值
值域：false、true
数据填充约束：M
备注：“false”表示数据为非生成数据，“true”表示数据为生成数据。

7 数据内容元数据

7.1 模态类型

中文名称：模态类型
英文名称：media_type
定义：数据的模态类型
数据类型：数组
值域：text、image、video、audio 等
数据填充约束：M
备注：此处列举的是数据的常见模态类型。当数据的模态类型不在列举范围内时，可采用符合实际的其他模态类型表示。

7.2 内容

中文名称：内容
英文名称：content
定义：数据的具体内容
数据类型：字符串
值域：自由文本
数据填充约束：M
备注：文本数据可用数据本身表示（宜使用 UTF-8 编码），图像、视频、音频等其他类型数据可用相对存储路径表示。

8 标注信息元数据

8.1 标签

中文名称：标签

英文名称：label
定义：数据的标签
数据类型：数组
值域：自由文本
数据填充约束：M
备注：具体内容根据数据集所针对的人工智能任务做进一步规定。若数据集的目标人工智能任务为无监督学习任务，数据填充约束为“0”。

8.2 标注方式

中文名称：标注方式
英文名称：annotation_method
定义：数据标注的方式
数据类型：字符串
值域：人工标注、自动标注、半自动标注、其他
数据填充约束：O
备注：无。

8.3 标注工具

中文名称：标注工具
英文名称：annotation_tool
定义：数据标注所使用的工具
数据类型：字符串
值域：自由文本
数据填充约束：O
备注：无。

8.4 标注人员类型

中文名称：标注人员类型
英文名称：annotator
定义：标注数据的人员类型
数据类型：字符串
值域：普通标注员、专业标注员、行业领域专家、其他
数据填充约束：O
备注：无。

附 录 A
（资料性）
数据集格式示例


```
{
    "id": "75110000050001381XW1100HN7FYS9X7",
    "rid": ["75110000050001381XW11000802X8KI5"],
    "data_content":[
        {
            "media_type": ["image"],
            "content": "../data/images/streetscape.jpg"
        }
    ],
    "annotation":{
        "label":[
            {
                "iscrowd": 0,
                "bbox": [20, 20, 20, 20],
                "category": "human"
            },
            {
                "iscrowd": 0,
                "bbox": [40, 40, 40, 40],
                "category": "car"
            }
        ],
        "annotation_method": "人工标注",
        "annotation_tool": "LabelImg",
        "annotator": "普通标注人员"
    },
    "original_time": "2025-01-01",
    "last_modified_time": "2025-01-01",
    "version": "1.0.0-alpha",
    "license": "其他",
    "source": "组织机构",
    "source_details": "中国电子工业标准化技术协会",
    "generated_data_indicator": false
}
```

参 考 文 献

[1] 20255407-T-907 数据 基础术语

[2] GB/T 18391.1-2009 信息技术 元数据注册系统（MDR） 第1部分：框架

[3] ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

[4] ECMA-404 The JSON data interchange syntax

[5] IETF RFC 8259 The JavaScript Object Notation (JSON) Data Interchange Format

[6] 语义化版本（Semantic Versioning） https://semver.org/lang/zh-CN