ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T XXXXX—XXXX

数据基础设施 数据目录描述要求

Data infrastructure—Data catalog description requirements

(点击此处添加与国际标准一致性程度的标识)

（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/TC609）提出并归口。

本文件起草单位：北京物资学院、北京化工大学、中国移动通信有限公司研究院、北京交通大学、浙江蚂蚁密算科技有限公司、上海数据交易所有限公司、杭州安恒信息技术股份有限公司、北京市大数据中心、上海零数众合信息科技有限公司、联通数据智能有限公司、中国电信集团有限公司、三六零数字安全科技集团有限公司、北京易华录信息技术股份有限公司、蚂蚁科技集团股份有限公司、华控清交信息科技（北京）有限公司、北京华宇软件股份有限公司、中国移动通信集团有限公司、成都数据集团股份有限公司、杭州金智塔科技有限公司、中国交通通信信息中心、中国电子科技集团公司第十五研究所、四川数通智汇数据科技有限公司、农业农村部大数据发展中心、北京腾云天下科技有限公司、南湖实验室、北京新材道数智科技有限公司、阿里巴巴（中国）有限公司、中航信数智科技（北京）有限公司、北京邮电大学、郑州数据交易中心有限公司、苏州数据资产运营有限公司、下一代互联网关键技术和评测北京市工程研究中心有限公司、杭州锘崴信息科技有限公司、中电云计算技术有限公司、云联智高（北京）信息技术研究院有限公司、湖州市数字集团有限公司、云宏信息科技股份有限公司、蓝象智联（杭州）科技有限公司、数据易（北京）信息技术有限公司。

本文件主要起草人：张茜茜、张闯、喻炜、孙祥栋、涂群、李征、王萍、刘世峰、宫大庆、徐广姝、韦韬、张晓蒙、刘圣威、于百程、陶立峰、周俊、林峰璞、兰春嘉、杨珍、宋雨伦、张鑫、胡振泉、耿贵宁、程宏、昌文婷、靳晨、魏丽丽、茹志强、邓建平、李梦杰、王林、曲薇、张冰、张文馨、张亚东、张磊、李慧玲、王琳、叶可、何帅、王平凡、杨红、张旭东、李帜、魏涛、白超、王震、朱敏健、王超、宾军志、刘锐剑、李大中、李锋、李杰、彭晋、隗樊、罗辰雨、李冠洲、王斌、李岩、赵娜、李由、卫炜、黄洋成、王畅畅、何运昌、孙晓峰、牛一锋、孙琪、方正、李旭、刘思佳、徐震兴、陆晓伟。

数据基础设施 数据目录描述要求

1 范围

本文件规定了数据基础设施中数据资源和数据产品目录的描述要求、技术要求和安全要求。

本文件适用于数场、可信数据空间、数联网、数据元件、隐私保护计算、区块链等技术体系支撑的各类层级的数据基础设施建设，包括区域、城市、行业、企业、个人等数据基础设施。

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 2260 中华人民共和国行政区划代码

GB/T 4754—2017 国民经济行业分类

GB/T 36344—2018 信息技术 数据质量评价指标

GB/T 35273 信息安全技术 个人信息安全规范

GB/T 43697 数据安全技术 数据分类分级规则

GB/T XXXXX—XXXX 数据基础设施 互联互通基本要求

GB/T XXXXX—XXXX 数据基础设施 标识要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据资源 data resource

具有价值创造潜力的数据的总称，通常指以电子化形式记录和保存、可机器读取、可供社会化再利用的数据集合。

[来源：20255407-T-907，3.1.4]

3.2

数据产品 data product

自然人、法人或者非法人组织对其合法获取的数据资源，经过实质性加工和创新性劳动后形成的，可满足特定需求的数据加工品和数据服务。

[来源：20255407-T-907，3.1.14]

3.3

数据目录 data catalog

用于描述数据特征的一组信息，以提高数据的可发现性、可理解性和可管理性。本文件中的数据目录包括数据资源目录和数据产品目录。

［来源：20255407-T-907，3.3.25］

3.4

数据资源目录 data resource catalog

用于分类、检索、定位数据资源的一组信息描述，包括但不限于数据资源名称、归属行业等数据资源的特征信息。

3.5

数据产品目录 data product catalog

用于分类、检索、定位数据产品的一组信息描述，包括但不限于数据产品名称、归属行业、交付方式等数据产品的特征信息。

3.6

高质量数据集 high-quality dataset

在完整性、准确性、一致性、及时性、有效性等质量指标上均达到优良水平，具备完善的数据标注、清晰的数据结构和良好的可追溯性，能够有效支撑人工智能模型训练、推理及其他数据应用场景的数据集合。

4 缩略语

下列缩略语适用于本文件。

API：应用程序编程接口（Application Programming Interface）

DOA：数字对象架构（Digital Object Architecture）

5 数据目录分类

5.1 数据资源目录分类

数据资源目录分类维度包括所属行业和数据来源，具体见表1。

表1 数据资源目录分类

分类维度	序号	资源分类	说明
数据资源所属行业	1-20	某行业数据资源	数据资源所属行业，依据GB/T 4754—2017分为20个大类。即农、林、牧、渔业；采矿业；制造业；电力、热力、燃气及水生产和供应业；建筑业；批发和零售业；交通运输、仓储和邮政业；住宿和餐饮业；信息传输、软件和信息技术服务业；金融业；房地产业；租赁和商务服务业；科学研究和技术服务业；水利、环境和公共设施管理业；居民服务、修理和其他服务业；教育；卫生和社会工作；文化、体育和娱乐业；公共管理、社会保障和社会组织；国际组织；可根据实际需求扩展细分行业类别，扩展代码从21开始编号。
数据资源来源	1	原始取得	数据主体通过自身的行为或活动直接创造出数据，从而获得对该数据的数据资源持有权、数据加工使用权。
	2	收集取得	数据主体通过各种方式从一个或多个来源处汇集相关数据资源。
	3	交易取得	数据主体以支付一定的对价（如货币、股权、其他数据资源等）

			为条件，获得数据资源持有权、数据加工使用权。
	4	共享取得	数据主体通过政府公共数据开放、行业联盟共享、企业间数据合作等方式获得的数据资源。
	5	其他	除上述来源渠道以外的其他方式，包括数据回流（上级系统向下级单位回流的数据）、数据授权使用等获得的数据资源。

5.2 数据产品目录分类

数据产品目录分类维度包括产品交付形式和所属行业，具体见表2。

表2 数据产品目录分类

分类维度	序号	产品分类	说明
数据产品交付形式	1	数据集	经过收集、整理和处理的结构化或非结构化的数据集合，可以是数字、文本、图像、音频、视频等类型的数据，通常用于分析、研究或支持决策。例如公共数据集、商业数据集、行业专用数据集等。
	2	API产品	通过API进行传输和交互的数据，利用标准化接口简化数据的获取和使用过程。例如公共API数据、商业API数据等。
	3	数据应用	数据资源经过软件、算法等技术手段处理后，可面向数据使用方提供数据呈现或操作的数据服务。例如数据处理云服务、数据分析软件等。
	4	数据报告	基于数据分析和处理生成的文档或展示材料，用于向特定的受众传达数据所蕴含的意义、趋势、关系以及结论。例如商业数据报告、财务数据报告、政府数据报告等。
	5	数字对象	对交付内容进行标准化封装的数字化产品形式，例如DOA数字对象。数据元件、数据件、算法模型等均可封装成数字对象。
	6	其他	除上述形态之外的数据产品形式，通常针对特定需求或应用场景，提供定制化和专业化的功能与服务。例如数据模型服务、数据标注工具、数据可视化服务等。
数据产品所属行业	1-20	某行业数据产品	数据产品所属行业，依据GB/T 4754—2017分为20个大类。即农、林、牧、渔业；采矿业；制造业；电力、热力、燃气及水生产和供应业；建筑业；批发和零售业；交通运输、仓储和邮政业；住宿和餐饮业；信息传输、软件和信息技术服务业；金融业；房地产业；租赁和商务服务业；科学研究和技术服务业；水利、环境和公共设施管理业；居民服务、修理和其他服务业；教育；卫生和社会工作；文化、体育和娱乐业；公共管理、社会保障和社会组织；国际组织；可根据实际需求扩展细分行业类别，扩展代码从21开始编号。

6 数据目录元数据

6.1 数据资源目录元数据描述规则

数据资源目录元数据描述规则见表3，数据资源目录的元数据项属性包括中文名称、字段名称、约束类型、数据类型、数据格式、取值说明等内容见附录A。

表3 数据资源目录元数据描述规则

中文名称	描述
资源名称	描述数据资源具体内容的标题名称。
数据资源标识码	全域唯一的数据资源标识。
行业分类	数据资源所属的国民经济行业的行业名称。依据GB/T 4754—2017定义的类别名称。
资源持有方	数据资源持有主体的名称。
资源持有方标识码	数据资源持有主体标识。
联系人	资源持有方联系人名称。
联系方式	资源持有方联系人的联系方式。
资源摘要	对数据资源内容（或关键字段）的概要描述，以3-5个短语为宜，以“；”隔开。
资源格式	数据资源的存在方式，数据资源持有方应提供可机读的电子格式及相关软件版本信息。
信息项名称	描述电子表格、数据库等结构化数据资源中具体数据项（字段）的中文标题。
信息项数据类型	标明该信息项的数据类型。
数据来源	获取该数据资源的方式，包括但不限于： 1)原始取得。数据主体通过自身的行为或活动直接创造出数据，从而获得对该数据的资源持有权、加工使用权或产品经营权； 2)收集取得。数据主体通过各种方式从多个分散的来源处汇集相关数据的过程； 3)交易取得。数据主体以支付一定的对价（如货币、股权、其他数据资源等）为条件，获得数据资源的持有权、加工使用权、产品经营权或其他相关权益。 4)共享取得。数据主体通过政府公共数据开放、行业联盟共享、企业间数据合作等方式获得的数据资源。 5)其他。数据主体通过除上述来源渠道以外的其他方式。
数据资源规模	数据资源的存储体量或规模范围，可按覆盖范围、存储空间、记录条数等方式表达。
覆盖时间范围	数据资源所包含的历史时间区间或时间跨度，如数据采集的时间范围。
数据资源更新频率	数据资源的更新频率，单位为次/天、次/周、次/月、次/季度、次/年、实时或不更新等。
数据更新方式	数据资源内容的更新机制，可分为全量更新（对全部数据进行整体替换）和增量更新（仅对新增、修改或删除部分进行更新）。
数据资源版本号	数据资源的版本标识，如V1.0、V2.0
是否涉及个人信息	按照数据资源是否涉及个人信息进行分类。依据GB/T 35273对个人信息（包含个人隐私数据）的定义和规范，方法见附录B。
数据资源状态	数据资源的当前状态（待登记、已登记、已撤销等）。
数据资源质量	依据GB/T 36344—2018对数据资源在完整性、准确性、一致性、及时性、有效性等方面的质量水平进行分类并说明。其中，高质量数据集应在上述各项指标上均达到优良水平，并具备完善的数据标注、清晰的数据结构和良好的可追溯性，能够有效支撑人工智能模型训练与推理等应用场景。
高质量数据集标识	标明该数据资源是否经过高质量数据集认定。取值包括： 1)已认定。经权威机构或第三方评估确认符合高质量数据集标准； 2)自评达标。资源持有方自行评估认为符合高质量标准；

	3)未评估。尚未开展高质量认定。
数据安全分级分类	依据GB/T 43697对数据资源涉及的数据安全进行分类并说明。
AI可用性等级	数据资源对人工智能应用的适用程度。取值包括：1)可直接用于训练；2)需预处理后可用；3)仅可用于推理；4)不适用于AI。
AI任务类型	数据资源适用的AI任务类型。包括但不限于：自然语言处理、计算机视觉、语音识别、推荐系统、时序预测、多模态大模型、强化学习等。
数据标注信息	数据资源的标注状态信息。包括：是否已标注、标注类型（分类标注、序列标注、回归标注、生成式标注等）、标注规模、标注质量指标等。
数据模态	数据资源的模态类型。包括：文本、图像、音频、视频、结构化数据、多模态等。对于多模态数据应说明模态组合方式及对齐关系。
是否含合成数据	标明数据资源中是否包含由AI模型生成的合成数据。若包含，应说明合成比例、合成方法及其与真实数据的关系。
AI训练授权状态	数据资源是否允许用于AI模型训练。取值包括：允许商用训练、仅允许科研训练、禁止用于训练、需单独授权。
其他	其他相关说明。

6.2 数据产品目录元数据描述规则

6.2.1 产品登记信息元数据

6.2.1.1 基本信息元数据

数据产品登记基本信息元数据描述规则见表4，数据产品登记基本信息元数据项属性包括中文名称、字段名称、约束类型、数据类型、数据格式、取值说明等内容见附录A。

表4 数据产品登记基本信息元数据描述规则

中文名称	描述
产品名称	描述数据产品具体内容的标题名称。
产品标识码	空，待登记完由系统自动分配。
产品类型	数据产品的类型。包括数据集、API产品、数据应用、数据报告、数字对象、其他。
覆盖时间范围	数据产品所包含的时间跨度，如数据集产品中数据采集的时间范围。
行业分类	数据产品所属的国民经济行业的行业名称。依据GB/T 4754—2017的类别名称。
地域分类	数据产品所属地域名称和代码。依据GB/T 2260的名称和数字码。
是否涉及个人信息	按照数据产品是否涉及个人信息进行分类，依据GB/T 35273对个人信息（包含个人隐私数据）的定义和规范，方法见附录B。
产品简介	数据产品的简要描述信息，包括但不限于数据产品的用途、内容、规模等。
交付方式	数据产品进行交付的方式。如文件传输、数据流传输、API传输等。
使用限制	数据产品在使用过程中的限制条件和范围。如产品不可进行二次加工或转卖等。
授权使用	数据产品使用时是否需要被关联的数据资源持有方授权。
数据主体	关联数据资源所涉及的数据主体，如个人信息、企业数据或公共数据所属主体。
数据规模	关联数据资源的体量或数据产品本身的体量，单位为MB、GB、TB等。
更新频率	关联数据资源的更新频率，单位为次/天、次/周、次/月、次/季度、次/年、实时或不更新等。

数据资源标识码	数据产品形成所依据的数据资源的标识，可能为唯一的数据资源标识，也可为一组数据资源的标识列表。
适用模型类型	数据产品适用的AI模型类型。包括：大语言模型（LLM）、多模态大模型、图像生成模型、语音模型、推荐模型、专用模型等。
AI用途场景	数据产品的AI应用场景。包括：预训练、指令微调（SFT）、RLHF对齐、检索增强生成（RAG）、知识库构建、评测基准、合成数据生成、Agent工具调用等。
其他	其他相关说明。

6.2.1.2提供方信息元数据

数据产品提供方信息元数据描述规则见表5，数据产品提供方信息元数据项属性包括中文名称、字段名称、约束类型、数据类型、数据格式、取值说明等内容见附录A。

表5 数据产品提供方信息元数据描述规则

中文名称	描述
提供方名称	数据产品提供方名称。
提供方主体类型	自然人、法人或非法人组织。
主体信息	提供方主体的基本信息。
身份标识码	提供方主体唯一身份标识。
提供方简介	提供方的情况简介，包括但不限于介绍提供方概况、业务范围、数据资源、相关资质等信息。
法人经办人姓名	企业或机构组织，经由法定代表人授权，负责执行数据产品发布等相关具体事务的个人姓名。
法人经办人电话	法人经办人的电话号码
法人经办人身份证	法人经办人的身份证信息。
授权委托书	企业或机构组织，法定代表人授权委托他人负责执行数据产品发布等相关具体事务的授权证明文件。

6.2.1.3声明信息元数据

数据产品声明信息元数据描述规则见表6，数据产品声明信息元数据项属性包括中文名称、字段名称、约束类型、数据类型、数据格式、取值说明等内容见附录A。

表6 数据产品声明信息元数据描述规则

中文名称	描述
数据样例	数据产品提供方在登记数据产品时可提交数据样例，即从数据产品所涉及的数据集中抽取的具有代表性的一部分数据。用于展示数据的结构、内容和特征，帮助用户快速了解数据的基本情况。
合法合规声明	数据产品提供方在登记数据产品时应提交数据产品合法合规声明。声明该数据产品在采集、开发过程中符合相关法律法规。声明模板见“附录 D 合法合规声明”。
数据来源声明	数据产品提供方在登记数据产品时应提交数据来源声明。证明加工该数据产品的数据来源有据可查，一般通过数据资产登记证书声明。声明模板见“附录 E 数据来源声明”。
安全分级分类	依据GB/T 43697对数据产品涉及的数据进行分类并说明。
数据质量、产品价值评估报告	数据产品提供方在登记数据产品时，若有数据质量检测报告、数据产品价值评估报告，可将其作为附件提供，用于明确产品价值、增加用户对数据产品的信任度。

数据标注规范	数据产品的标注规范信息。包括标注体系说明、标注工具、标注人员资质、质量控制流程（如多人标注一致性检验、Kappa系数等）。
合成数据声明	若数据产品包含合成数据，应声明合成方法（如使用的模型名称、版本）、合成比例、与真实数据的差异性评伊结果、是否存在数据污染风险等。
AI训练授权许可	数据产品用于AI训练的授权许可信息。包括许可类型（开源协议、商用许可、科研专用等）、许可范围、模型输出物的权利归属等。

6.2.2 产品上架信息元数据

产品上架时需要提供的信息，包括上架基本信息、定价信息、交付方式信息，产品上架后根据数据产品标识码关联数据产品登记平台信息。

6.2.2.1上架基本信息元数据

数据产品上架基本信息元数据描述规则见表7，数据产品上架基本信息元数据属性包括中文名称、字段名称、约束类型、数据类型、数据格式、取值说明等内容见附录A。

表7 数据产品上架基本信息元数据描述规则

中文名称	描述
数据产品标识码	数据产品登记完成后唯一标识。
上架业务节点名称	数据产品发布的业务节点名称。
上架业务节点ID	数据产品发布的业务节点ID。
上架业务节点位置	数据产品发布到业务节点上的具体位置信息，例如数据产品链接。
上架交付方式	数据产品在当前业务节点上架时允许的交付或联合加工方式。应基于产品登记时的交付方式，结合数据安全分级分类要求进行限定。包括：文件交付、API交付、数据流交付、沙箱计算、隐私计算、数据可视化、其他。涉及敏感个人信息或核心/重要数据的，宜优先选择沙箱计算或隐私计算方式。
其他	其他相关说明。

6.2.2.2定价信息

数据产品应按次数、按周期、一事一议、其他等方式进行定价或可根据实际情况选择免费提供，具体定价内容描述信息请参考附录C。

6.2.2.3交付方式信息

数据产品应提供数据产品交付方式说明，具体见表8，明确访问和使用数据的要求，数据产品交付方式的属性包括中文名称、字段名称、约束类型、数据类型、数据格式、取值说明等内容见附录A。**表**

表8 数据产品交付方式说明

交付方式	要求
文件交付	1) 应明确文件交付形式，如平台下载、文件传输服务等，若为文件传输服务，应明确文件传输协议，如HTTP、HTTPS、FTP等； 2) 应明确数据文件存储的编码方式，如UTF-8等； 3) 应明确传输过程中的文件是否加密，若加密，需说明加密方式； 4) 应明确文件交付地址信息，如接入连接器ID、交付地址等。

数据流交付	1) 应明确数据流传输协议，如WebSocket、MQTT、Apache Kafka、gRPC等； 2) 应明确数据流的格式，如JSON、Avro、Protobuf、XML等； 3) 应明确数据流的编码方式，如UTF-8、二进制编码等； 4) 应明确传输过程中的数据是否加密，若加密，需说明加密方式； 5) 应明确数据流交付地址信息，如接入连接器ID、交付地址； 6) 应明确数据流传输方式，包括流式传输（实时连续传输）和批式传输（定时批量传输）。
API交付	1) 应明确API服务传输协议，如HTTP、HTTPS等； 2) 应明确规定API服务所使用的接口类型，如 RESTful API、Webservice等； 3) 应明确API输入、输出参数及相关定义； 4) 应明确定义数据传输格式，如JSON、XML等； 5) 应明确API服务的请求数据包、响应数据包体格式； 6) 应提供API服务响应代码及描述说明信息； 7) 应提供API请求、响应示例说明； 8) 可补充附件文件，作为API服务对接说明； 9) 应明确API交付地址信息，如接入连接器ID、交付地址。
其他交付方式	包括定制化服务、咨询服务、数据加工服务等非标准化交付形式，应在交付方式说明中详细描述具体的交付内容、方式和要求。

7 数据目录技术要求

7.1 目录编制要求

用户编制待登记的数据目录，应对照附录A数据目录字典，确保字段值符合字典约束，保证数据目录的完整性、规范性、准确性和一致性，避免遗漏或错误描述。可引入人工智能技术辅助目录编制，支持智能化预填充和建议。

7.2 目录传输要求

数据基础设施体系下，各节点间传输数据目录，应遵循以下要求：

a) 应实现数据目录实时传输，确保数据在不同节点间快速同步；
b) 应基于标准化接口实现目录传输（依据《GB/T XXXXX—XXXX 数据基础设施 互联互通基本要求》），确保不同节点之间的互联互通；
c) 应提供传输状态实时监控功能，包括传输成功率、响应时间、错误率、网络延迟等关键指标；
d) 应支持大文件目录的断点续传和完整性校验机制。
e) 应支持数据目录的语义化描述和机器可读格式（如JSON-LD、RDF等）传输，以支持AI系统对数据目录的自动解析和理解

7.3 目录管理要求

对归集的数据目录进行管理，应遵循以下要求：

a) 应说明数据目录的上报、修改以及下架等流程，以便用户依照合规流程开展操作，针对不符合既定流程要求的操作，予以驳回处理；
b) 应支持对上报的数据目录内容进行初步审核，可引入人工智能技术辅助，检查其完整性、规范性、准确性和一致性，通过格式审核的目录，经人工审核确认后自动接收；

c) 应具备数据目录的存储、备份与恢复功能，以保障数据目录的安全性和可用性；
d) 应对数据目录进行统一维护，包括按要求更新数据目录，以及对其建立多维度分类标签等；
e) 可支持基于身份的访问控制与动态授权；
f) 应支持目录快照和历史版本回溯。
g) 应支持基于AI的智能目录治理功能，包括利用自然语言处理技术自动识别和标记重复、冲突或低质量的目录条目；利用模型识别数据目录之间的关联关系，构建数据目录知识图谱等。

7.4 目录查询要求

数据基础设施体系下，提供数据目录查询功能的节点，向用户提供数据目录查询服务。应遵循以下要求：

a) 应支持单个检索词检索、多个检索词组合检索等多种检索方式，多关键词检索应明确逻辑关系，默认采用“与”逻辑组合；
b) 可采用分类导航的模式，便于用户快速定位和查询；
c) 应引入人工智能技术赋能目录查询，辅助进行语义检索、跨模态检索、关联查询、智能推荐、快速定位检索等系列工作，使用AI技术时应确保不访问未授权的原始数据，仅基于公开的元数据进行智能推理。

8 数据目录安全要求

8.1 管理安全要求

a) 数据目录的内容，以及编制、管理和应用的过程应严格遵守国家法律法规和相关政策要求，如《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等；
b) 应构建数据目录责任制度，详细规定各环节的操作规范与安全准则，明确数据安全责任主体，并清晰责任分工，建立追责问责机制；
c) 应对数据目录中提交样例数据部分涉及个人隐私的数据进行脱敏处理，确保数据目录中不泄露个人隐私；
d) 应根据不同安全等级的数据，数据资源管理责任主体对数据目录及其对应数据样例设置不同的公开范围，拟公开的信息，应未涉及任何国家安全与秘密的信息；
e) 应定期进行数据目录安全评估和风险监测，确保数据目录的安全性。
f) 向用户提供目录查询服务时，应对用户身份进行鉴别和认证，确保用户访问权限与数据目录安全等级相匹配。

8.2 技术安全要求

a) 数据基础设施体系中，数据目录传输所采用的设备应符合有关设备的安全要求；
b) 应对已限制访问的数据目录，设置访问控制体系，防止未经授权的访问和信息泄露。

附 录 A
（规范性）
数据目录字典

数据目录字典涉及的数据类型应符合以下类型：

a) 字符型：描述字符类型的属性，字母简称为C；
b) 数值型：描述整数、浮点数等类型的属性，字母简称为N；
c) 布尔型：描述是/否、真/假等类型的属性，字母简称为B；
d) 日期型：描述有日期相关的属性；
e) 二进制型：描述文件等类型的属性；
f) 对象型：描述一个主体的信息集合；
g) 数组型：描述一组同类信息的集合。

数据目录字典的数据格式应符合以下类型：

a) 固定长度：在字母C后直接给出字符长度的数目，如C4表示长度为4的字符串；
b) 可变长度：在字母C后加“..”，再给出接口字段最大字符数目，如C..4，表示最大长度4位的字符串。

数据资源元数据描述见表A.1，数据产品元数据描述见表A.2。

表A.1 数据资源目录字典

中文名称	字段名称	约束类型	数据类型	数据格式	取值说明
1.资源名称	resourceName	必填	字符型	C..128	无
2.数据资源标识码	resouceId	必填	字符型		依据《GB/T XXXXX—XXXX 数据基础设施 标识要求》的编码规则
3.行业分类	industry	必填	字符型	C1	依据GB/T 4754—2017的门类代码
4.资源持有方	resourceOwner	必填	字符型	C..128	无
5.资源持有方标识码	OwnerId	必填	字符型		依据《GB/T XXXXX—XXXX 数据基础设施 标识要求》的编码规则
6.联系人	contacter	必填	字符型	C..10	无
7.联系方式	contactInformation	必填	字符型	C11	无
8.资源摘要	resoureceAbstract	必填	字符型	C..1024	无
9.资源格式	resourceFormat	必填	字符型	C..24	电子文件存储格式：OFD、wps、xml、txt、doc、docx、html、pdf、ppt等； 电子表格存储格式：et、xls、xlsx等； 数据库类存储格式：Dm、KingbaseES、access、dbf、dbase、sysbase、oracle、sql server、db2等； 图形图像类存储格式：jpg、gif、bmp等； 流媒体类存储格式：swf、rm、mpg等； 自描述格式：由持有方提出其特殊行业领域的通用格式。
10.信息项名称	itemName	选填	字符型	C..128	无
11.信息项数据类型	itemType	选填	字符型	C..128	文本类信息，应标明所采用的字符集和编码方式； 结构化数据，应标明数据类型及数据长度。
12.数据来源	dataSource	必填	字符型	C2	01：原始取得 02：收集取得 03：交易取得 04：共享取得 05：其他
13.数据资源规模	dataSize	选填	字符型	C..128	01：按记录条数表示，单位为条 02：按存储空间表示，单位为MB、GB、TB等 03：按覆盖范围表示，适用于空间类、业务类资源 04：混合表示，适用于当资源类型复杂，需结合多种方式共同描述

表A.1 数据资源目录字典（续）

中文名称	字段名称	约束类型	数据类型	数据格式	取值说明
14.覆盖时间范围	timeRange	选填	日期型	YYYY-MM-DD至YYYY-MM-DD	无
15.更新频率	updateFreqency	必填	字符型	C..10	单位为次/天、次/周、次/月、次/季度、次/年、不更新、其他
16.数据更新方式	updateMethod	选填	字符型	C..10	全量、增量
17.数据资源版本号	versionNo	选填	字符型	C..10	数据资源的版本标识，如V1.0、V2.0
18.是否涉及个人信息	personalInformation	必填	字符型	C1	0：不涉及 1：一般个人信息 2：敏感个人信息
19.资源状态	resourceStatus	必填	字符型	C2	01：待登记 02：已登记 03：已撤销
20.数据资源质量	resourceQuality	选填	字符型	C2	01：高质量（满足准确性、完整性、一致性、及时性、有效性要求，且具备支撑AI训练等高标准应用的数据质量水平） 02：一般质量（部分质量指标存在轻微偏差，对业务影响有限） 03：低质量（多个指标不达标，需修复或清洗） 04：未评估（尚未开展质量评估）
高质量数据集标识	hqDataset Flag	选填	字符型	C2	01：已认定 02：自评达标 03：未评估
21.数据安全分级分类	safeLevel	选填	字符型	C2	01：核心数据 02：重要数据 03：一般数据
22.AI可用性等级	aiReadiness	选填	字符型	C2	01:可直接训练 02:需预处理 03:仅可推理 04:不适用AI
23.AI任务类型	aiTaskType	选填	数组型		NLP、CV、ASR、推荐、时序、多模态、RL、其他
24.数据模态	dataModality	选填	数组型		文本、图像、音频、视频、结构化、多模态
25.数据标注信息	annotationInfo	选填	对象型		包括标注状态、标注类型、标注规模、质量指标

26.是否含合成数据	hasSyntheticData	选填	字符型	C2	01:不包含 02:部分包含 03:全部合成
27.AI训练授权	aiTrainingAuth	选填	字符型	C2	01:允许商用 02:仅科研 03:禁止训练 04:需单独授权
28.其他	others	选填	对象型		可扩展信息

表A.2 数据产品目录字典（产品登记信息）

类别	中文名称	字段名称	约束类型	数据类型	数据格式	取值说明
1.基本信息	1.1产品名称	productName	必填	字符型	C..128	无
	1.2产品标识码	productId	选填	字符型		空，待登记完由系统自动分配。
	1.3产品类型	productType	必填	字符型	C2	01：数据集 02：API产品 03：数据应用 04：数据报告 05：数字对象 06：其他
	1.4覆盖时间范围	timeRange	选填	日期型	YYYY-MM-DD至YYYY-MM-DD	无
	1.5行业分类	industry	必填	字符型	C1	依据GB/T 4754—2017的门类代码
	1.6地域分类	productRegion	选填	字符型	C..12	依据GB/T 2260的名称和数字码
	1.7是否涉及个人信息	personalInformation	必填	字符型	C1	0：不涉及 1：一般个人信息 2：敏感个人信息
	1.8产品简介	description	必填	字符型	C..500	无
	1.9交付方式	deliveryMethod	必填	字符型	C2	01：文件传输 02：数据流传输 03：API传输 04：其他
	1.10使用限制	limitations	必填	字符型	C..2048	无
	1.11授权使用	authorize	必填	布尔型	B	0：否 1：是
	1.12数据主体	dataSubject	必填	字符型	C2	01：个人信息 02：企业数据 03：公共数据
	1.13数据规模	dataSize	选填	数值型	N16	单位为MB、GB、TB等
	1.14更新频率	updateFreqency	必填	字符型	C..10	单位为次/天、次/周、次/月、次/季度、次/年、其他
	1.15数据资源标识	resourceId	选填	数组		一个或多个数据资源标

	码			型		识码
	1.16适用模型类型	applicableModelType	选填	数组型		数据产品适用的AI模型类型。取值包括： 01：大语言模型（LLM） 02：多模态大模型 03：图像生成模型 04：语音模型 05：推荐模型 06：专用模型 07：Agent/智能体模型 08：其他
	1.17AI用途场景	aiUsageScenario	选填	数组型		数据产品的AI应用场景。取值包括： 01：预训练 02：指令微调 03：RLHF/RLAIF对齐 04：检索增强生成（RAG） 05：知识库构建 06：评测基准 07：合成数据生成 08：Agent工具调用 09：模型蒸馏/压缩 10：其他
	1.18其他	others	选填	对象型		可扩展信息
2．提供方信息	2.1提供方名称	providerName	必填	字符型	C..128	无
	2.2提供方主体类型	providerType	必填	字符型	C2	01:自然人 02：法人 03：非法人组织
	2.3 主体信息	entityInformation	必填	对象型		法人：注册地址、法定代表人、经营范围、成立日期、注册资本等能够明确法人身份和经营状况的详细内容； 非法人组织：社会组织类型、业务主管单位、登记管理机关、法定代表人、业务范围等信息； 自然人：性别、年龄、住址等信息。
	2.4身份标识码	identityId	必填	字符型		依据《GB/T XXXXX—XXXX 数据

						基础设施 标识要求》的编码规则
	2.5提供方简介	providerDesc	必填	字符型	C.1024	无
	2.6法人经办人姓名	operatorName	选填	字符型	C..10	无
	2.7法人经办人电话	operatorTelephone	选填	字符型	C11	无
	2.8法人经办人身份证	operatorIdCard	选填	字符型	C18	无
	2.9授权委托书	commission	选填	二进制型	/	形式为文件
3.声明信息	3.1数据样例	dataSample	选填	二进制型	/	形式为文件
	3.2合法合规声明	complianceAndLegal Statement	必填	二进制型	/	形式为文件
	3.3数据来源声明	dataSourceStatement	必填	二进制型	/	形式为文件
	3.4安全分级分类	safeLevel	选填	二进制型	/	形式为文件
	3.5数据质量、产品价值评估报告	evaluationReport	选填	二进制型	/	形式为文件
	3.6数据标注规范	annotationSpec	选填	二进制型	/	形式为文件
	3.7合成数据声明	syntheticDataDeclaration	选填	二进制型	/	形式为文件
	3.8AI训练授权许可	aiTrainingLicense	必填	二进制型	/	形式为文件

表A.3数据产品元数据描述（产品上架信息）

类别	中文名称	字段名称	约束类型	数据类型	数据格式	取值说明
1.上架基本信息	1.1数据产品标识码	productId	必填	字符型		依据《GB/T XXXXX—XXXX 数据基础设施 标识要求》的编码规则
	1.2上架业务节点名称	platformName	必填	字符型	C..128	由系统根据当前操作节点自动填充
	1.3上架业务节点ID	platformId	必填	字符型		依据《GB/T XXXXX—XXXX 数据基础设施 标识要求》的编码规则
	1.4上架业务节点位置	platformLocation	必填	字符型	C..128	无
	1.5上架交付方式	deliveryMode	选填	字符型	C2	01：文件传输 02：数据流传输 03：API传输 04：沙箱计算 05：隐私计算 06：数据可视化 07：其他
	1.6其他	others	选填	对象型		可扩展信息
2.定价信息	2.1计量方式	measureMethod	必填	字符型	C..6	参照附录C
	2.2计量单位	unit	必填	字符型	C..6	参照附录C
	2.3价格	price	必填	数值型	N8	无
3.交付信息	3.1交付方式说明	deliveryInfo	必填	二进制型	/	形式为文件

注：数据产品登记时，需填写基本信息、提供方信息和声明信息相关内容；数据产品上架时，需进一步完善上架基本信息、定价信息和交付信息相关内容

附 录 B
（规范性）
个人隐私保护说明

是否涉及个人信息分类编码见表B.1。

表B.1 是否涉及个人信息分类编码

编码	隐私分类	说明
0	不涉及个人信息	指不直接关联到个人身份或敏感性的一般信息。数据经过处理，使得无法直接识别特定个人，例如统计数据、去标识化的数据集以及在不涉及个人具体身份的情况下的业务数据。虽然这类数据的敏感度较低，但依然需要按照数据保护的原则进行合理管理和保护，以维护数据的整体安全。
1	一般个人信息	指能够单独或与其他信息结合识别特定自然人身份的信息，但泄露后对个人权益影响较小。例如姓名、出生日期、身份证号、联系方式、住址、工作单位等。
2	敏感个人信息	主要是指涉及个人隐私的敏感信息，包括但不限于个人生物识别信息、个人健康生理信息、个人财产信息、个人通信信息、个人位置信息等。这类信息一旦泄露或被滥用，可能对个人的隐私权造成直接损害。敏感个人信息的识别应参照《中华人民共和国个人信息保护法》第二十八条的相关规定。

附 录 C
（规范性）
数据产品定价信息描述示例

产品定价信息见表C.1。

表C.1 产品定价信息

产品类型	计量方式	计量单位	价格	计费说明
数据集	按用量	元/MB		按照数据集的数据传输容量来定价；
		元/GB		
		元/TB		
	按周期	元/天		基于单位时间段内的使用频率来定价；
		元/月		
		元/年		
	按条数	元/条		对于结构化数据，按数据条目来定价
API产品	按次数	元/次		根据接口被调用的次数计费；
	按订阅	元/月		按照订阅时长收费；
		元/年		
	按流量	元/MB		按照通过接口传输的数据量定价；
		元/GB		
		元/TB		
	按周期	元/天		基于更新的频率来定价；
		元/月		
		元/年		
数据应用或数据报告	按用量	元/MB		根据数据提供方提供的服务方式，定制化服务、分析报告等方式基于服务次数收费；应用服务调用、模型训练等服务模式基于用量收费。
		元/GB		
		元/TB		
	按次数	元/次		
	按订阅	元/月		按照订阅时长收费；
		元/年		
	按周期	元/天		基于更新的频率来定价；
		元/月		
		元/年		
其他	一事一议			根据具体数据产品内容、应用场景、使用频率和需求进行个性化的商议定价。

附 录 D
（资料性）
合法合规声明

致[数据使用方名称]：

感谢选择我司[数据产品名称]数据集/API产品/数据应用/数据报告。

针对向贵司提供的数据产品的合法合规情况，我司声明如下：

1.数据来源合法性

我司所提供的产品中涉及的数据内容，均来源于合法途径获取，包括但不限于合法自主收集/委托收集/外部采购/合作共享/授权运营等（一种或多种，结合产品实际），且已获得必要的授权或许可，不存在数据来源违法的情形。

2.数据处理合规性

我司所提供的产品中涉及的数据处理，均采用合规方式实施，包括但不限于对数据的收集、存储、使用、加工、传输、提供、公开、删除等，均遵循适用的法律法规、政策和监管要求，不存在数据处理违规的情形。

3.个人信息保护

我司采取了合理、必要的技术和组织措施，有效保护个人信息和隐私数据。

4.数据安全保障

我司建立了完善的数据安全管理体系，保障数据的安全性、完整性和可用性。

5.权益保护

我司提供的产品在合法合规前提下开展，遵守相关要求和合同约定，不存在对个人权益、企业权益、公众利益、社会秩序、经济运行、国家安全等造成威胁、影响、危害的情况。

6.知识产权

我司保证所提供的产品不侵犯任何第三方的知识产权，包括但不限于专利、商标、版权等。

我司将持续关注数据产品相关合法合规要求（包括但不限于法律、行政法规、地方性法规、部门规章、司法解释等）的变化，不断完善我们的技术和组织措施，接受相关有权部门、贵司及社会相关方的监督，确保数据产品满足合法合规要求。

特此声明。

[公司名称]

[声明日期]

附 录 E
（资料性）
数据来源声明

尊敬的客户：

感谢您选择我们的[产品名称]数据集/API产品/数据应用/数字对象/数据报告。在向您提供这些产品的过程中，

关于我们所使用数据的来源声明如下：

1.内部生成数据

部分数据是通过我们自身的业务运营和系统收集生成的，例如[具体业务活动或系统名称]。

这些数据的收集遵循了合法、公正和透明的原则，并采取了适当的技术和管理措施确保数据的准确性和完整性。

2.合作伙伴提供的数据

我们严格筛选评估合作伙伴，通过数据流通利用基础设施依法对[产品名称]数据集/API产品/数据应用/数字对象/数据报告涉及的数据合法持有。我们与合作伙伴签订了具有法律约束力的数据共享协议（数据许可协议/数据服务协议），明确了数据的使用目的、范围和安全保护措施，合作伙伴包括但不限于[列出主要合作伙伴的名称或类型]，以上合作伙伴提供的数据主要用于[说明数据的具体用途]。

3.公开数据

某些数据来源于合法的公开渠道，如政府公开数据、权威研究机构发布的数据等。

在获取和使用这些公开数据时，我们严格遵循了相关的使用规定和许可要求。

4.从个人信息主体处获取的数据

某些数据来自个人信息主体的，我们严格遵守《中华人民共和国个人信息保护法》等法律规定，从个人信息主体处获得相应的授权，不存在以欺诈、诱骗、误导等方式或从非法、违规渠道获取的数据。

我们郑重承诺，所使用的数据来源合法合规，并且在数据的处理、存储和使用过程中，始终严格遵守适用的法律法规和道德规范，不存在违反法律法规等的强制性规定或者危害国家安全、公共安全或侵犯第三方合法权益的情形，致力于保护用户的隐私和数据安全。

[公司名称]

[声明日期]

参 考 文 献

[1] GB/T XXXXX—XXXX 数据基础设施 参考架构
[2] GB/T XXXXX—XXXX 数据基础设施 用户身份管理和接入
[3] GB/T XXXXX—XXXX 数据基础设施 连接器技术要求