ICS 35.240
L 70

团 体 标 准

T/BSIA 00X-2026

衍生数据认定方法 第1部分 总则

Method standard for testing the identification of derivative data
part1 general principal

（征求意见稿）

2026-xx-xx 发布　　2026-xx-xx 实施

北京软件和信息服务业协会　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。

《衍生数据认定方法》分为5个部分：

——第1部分：总则；

——第2部分：通信行业；

——第3部分：矿山行业；

——第4部分：医疗健康行业；

——第5部分：情感认知行业。

本部分为《衍生数据认定方法》第1部分。

本文件由北京市大数据中心提出，由北京软件和信息服务业协会归口。

本文件起草单位：北京市大数据中心、智慧足迹数据科技有限公司、中国联合网络通信有限公司北京市分公司、北京认知洞察科技有限公司、北京健康有益科技有限公司、煤炭科学研究总院、北京软件和信息服务业协会。

本文件主要起草人：赵章界、赵莹、方方、施含章、赵华、吕振国、袁梦童、韩旭辉、吴爱琳、刘颖哲、李宇欣、裘实、李斯琦、李俊卿、骆意、杨培培、黄晴。

衍生数据认定方法 第1部分 总则

1 范围

本文件规定了衍生数据的认定规则、认定方法、实践使用、认定流程以及认定结果的表述要求。

本文件适用于对通过数据处理活动生成的数据是否属于衍生数据认定的各类场景。本文件所描述的衍生数据认定是通用的方法框架和指导性流程。特定行业、领域或特定类型的数据(如个人信息匿名化数据、加密数据等)的衍生数据认定，应在遵循本文件基本原则、流程和方法的基础上，结合相关法律法规、行业规范和技术特性，制定更细致的实施细则或补充判定标准。

本文件不涉及对原始数据、中间数据或已认定衍生数据的数据质量、数据安全技术细节(如加密、脱敏、匿名化效果)的详细评价和技术审计。

本文件为数据交易机构、数据第三方专业服务机构、行业主管部门、司法部门在在数据安全合规、数据资产治理、数据交易流通等场景中提供参考。

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 35295—2017 信息技术大数据术语

GB/T 36344—2018 信息技术数据质量评价指标

GB/T 38667—2020 信息技术大数据数据分类指南

3 术语和定义

下列术语和定义适用于本文件。

3.1

原始数据 Original Data

是指初次产生或源头收集的、未经加工处理的数据。

3.2

源数据 Source Data

数据处理者对其享有使用权，在保护各方合法权益前提下，用于生成衍生数据的全部初始数据。

注：源数据包括数据处理者声称的及潜在的数据。

3.3

中间数据 Intermediate Data

在数据处理过程中，由原始数据或其他中间数据经过部分处理产生，主要用于后续处理流程，不作为独立数据资产对外提供或长期存储的过程性数据。

注：中间数据可能临时存储于内存、缓存、临时文件或临时表中，其关键在于其目的和生命周期通常局限于特定的处理任务。

3.4

衍生数据 Derived Data

衍生数据，是指数据处理者对其享有使用权的数据，在保护各方合法权益前提下，通过利用专业知识加工、建模分析、关键信息提取等方式实现数据内容、形式、结构等实质改变，从而显著提升数据价值，形成的数据。

注：衍生数据应具有不可逆性，无法逆变换为源数据，且通常在形态、结构、价值或含义上与其输入数据存在显著差异（见3.9），或者代表了对输入数据的某种提炼、聚合、转化或洞察。

3.5

数据处理 Data Processing

包括数据的收集、存储、使用、加工、传输、提供、公开等。

3.6

认定 Identification

根据预设的标准、方法和流程，对特定数据是否满足衍生数据的定义和特征进行判断并给出结论的过程。

3.7

数据血缘关系 Data Lineage

描述数据从输入、处理到输出的全链路流转关系。

3.8

关键数据转换操作 Key Data Transformation Operations

在数据处理过程中，能够导致输入数据在形态、结构、内容或含义上发生实质性改变的操作类型。

3.9

显著差异 Significant Difference

指认定对象与输入数据在结构、内容或统计分布上存在可量化的根本性改变。

注：当其相对熵（见3.12）或信息熵变化率（见3.13）超过附录F中规定的推荐阈值时，应判定为存在显著差异。

3.10

重大变化 Major Change

指导致已认定数据性质可能发生根本性改变的处理逻辑或输入数据变化。

注：重大变化包括但不限于：处理逻辑中增减关键转换操作、输入源数据类型或主体发生根本改变、核心算法模型

（特别是生成新信息的模型）更换、导致结果数据关键指标计算方法或含义发生改变的调整。

3.11

信息熵 Information Entropy

衡量一个随机变量不确定性的数学度量。在本文件中，用于量化一个数据集内部状态的混乱程度或其所包含的信息量。数据集的熵值越高，其不确定性越大。具体定义与计算方法见附录F。

3.12

相对熵 Relative Entropy （Kullback-Leibler Divergence）

用于衡量两个概率分布之间差异的非对称性度量。在本文件中，用于量化衍生数据的概率分布相较于原始数据概率分布的改变程度。相对熵值越大，表明数据加工带来的改变越“实质性”。简称KL散度。具体定义与计算方法见附录F。

3.13

信息熵变化率 Entropy Change Rate (ECR)

衍生数据信息熵与原始数据信息熵之差的绝对值，与原始数据信息熵的比率。该指标直观反映了数据不确定性的相对变化幅度。具体定义与计算方法见附录F。

4 缩略语

API：Application Programming Interface（应用程序接口）

ETL：Extract，Transform，Load（数据抽取、转换、加载）

NLP：Natural Language Processing（自然语言处理）

AIGC：AI Generated Content (AI生成内容)

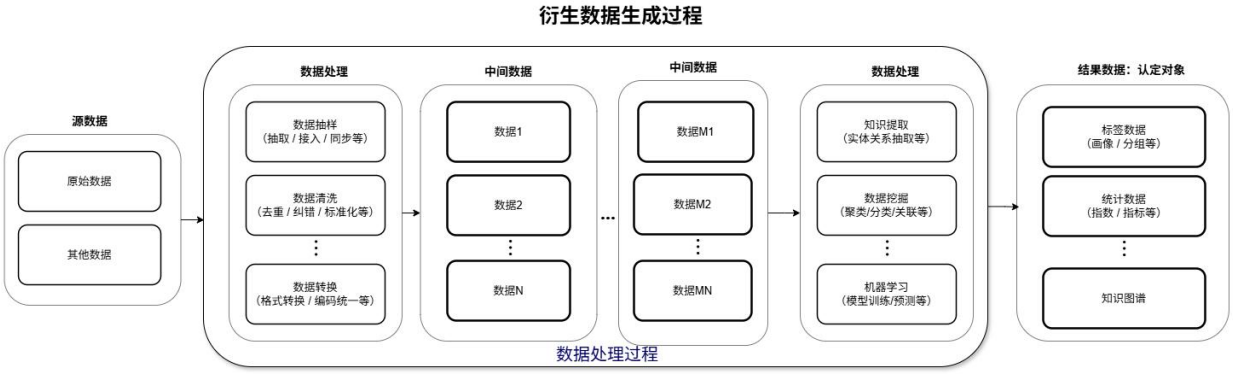
KL散度：Kullback-Leibler Divergence (相对熵)

ECR：Entropy Change Rate (信息熵变化率)

5 总则

5.1 衍生数据生成过程

认定衍生数据首先得明确源数据和认定对象，认定对象通常是由源数据通过一系列数据处理过程生成的结果数据，源数据包括原始数据和其他数据，数据处理过程中会生成若干中间数据，结果数据可能是衍生数据，也可能不是衍生数据。



衍生数据认定应当在明确结果数据对源数据依赖基础上，通过分析数据处理过程或比对结果数据与源数据的差异性，给予认定。

5.2 基本原则

衍生数据认定工作必须遵循以下基本原则：

a) 客观性：认定过程必须基于事实和证据，不应受主观判断或外部因素影响；

b) 可追溯性：认定结论应能够追溯到所依据的源数据及其处理过程、采用的认定方法及其结果；

c) 全面性：在界定的认定范围内，认定工作应尽可能覆盖数据处理的完整链条，分析关键环节，不应遗漏可能产生衍生数据的处理操作；

d) 审慎性：对于复杂或难以确定的情况，宜组合使用多种认定方法，并必须进行人工复核，确保认定的准确性和可靠性；

e) 处理过程优先：在判定过程中，处理过程分析（特别是对关键数据转换操作的识别）是基础性、优先性的判定依据。当处理过程分析和结果数据特征比对的初步结论冲突时，应以处理过程分析的结论为主，并加强人工复核。

5.3 基本要求

进行衍生数据认定工作时，应满足以下基本要求：

a) 明确认定对象与范围：在启动认定前，必须清晰界定本次需要认定的具体数据集或数据项的主键、边界和范围；

b) 获取完整必要信息：必须尽力获取与认定对象生成过程相关的全部必要信息，包括但不限于源数据及中间数据信息、完整的数据处理脚本/代码/配置、数据血缘信息、业务逻辑说明等。当无法获取完整信息时，应基于可获得的信息进行判定，并在认定报告中明确说明信息缺失情况及其对判定结论的潜在影响；

c) 规范选用认定方法：必须根据认定对象的特点、数据的复杂性、可获取的信息程度，规范选用本文件第 7 章规定的一种或多种认定方法，并优先进行处理过程分析；

d) 详细记录认定过程：必须完整、准确地记录认定过程中的所有关键信息，包括采用的方法、分析过程、依据的事实和证据、遇到的问题及解决方案、最终的判定结论等；

e) 保障认定过程独立性：认定人员或自动化认定系统宜独立于数据处理过程的开发、执行和日常运营团队，以确保判定结果的公正性；

f) 结果管理与更新：已完成的认定结果应妥善存储和管理。当认定对象的数据处理逻辑、源数据发生重大变化时，应重新进行认定或复核，并更新认定结果。

6 衍生数据的认定规则

6.1 数据来源关系

判断认定对象在内容或结构上是否逻辑地、技术地依赖于一个或多个其他数据作为输入。

6.2 数据处理过程

分析数据从输入到认定对象生成过程中发生的具体加工、分析、整合等活动，研判这些活动是否改变了数据的基本形态、结构、内容或生成了新含义，判定认定对象是否具有不可逆性且发生实质性改变。

6.3 结果数据特征

比对认定对象与源数据在结构、数据类型、统计特性、数据分布、内容相似度等方面的差异，判断这些差异是否达到显著的程度。此处的显著差异必须能够通过已识别的关键数据转换操作或复杂依赖关系得到合理解释。

7 认定方法

7.1 数据处理过程分析法

7.1.1 适用范围与前提条件

本方法适用于能够获取数据从输入到输出的详细处理逻辑、算法或代码的场景，是准确识别和评估是否存在关键数据转换操作的基础性方法。

前提条件：须能够访问或获取描述数据处理过程的清晰文档、脚本、代码或配置信息。

7.1.2 应获取到的关键素材

——数据处理流程图；

——系统设计文档；

——处理脚本/代码(如 SQL、Python、Java 等)；

——数据处理平台的 ETL 作业配置；

——算法模型定义文件；

——计算逻辑说明；

——日志文件（可选）。

7.1.3 处理过程

7.1.3.1 获取和解析处理信息

应获取与认定对象生成相关的完整数据处理脚本、代码或详细配置文档。应使用人工审查、代码分析工具或数据处理平台界面，解析并理解源数据、中间处理环节、逻辑和规则。

7.1.3.2 识别关键数据转换操作

应识别数据处理过程中所有涉及改变数据形态、结构、内容或生成新含义的关键操作。这些操作包括但不限于：

——多源或多表数据的复杂关联(JOIN)或合并(UNION)：非简单的基于单个字段的等值连接，而是涉及多字段、非等值、外部连接、或需要进行实体对齐后才能关联的操作；

——对数据进行分组聚合(GROUPBY)，计算统计量：计算涉及多个输入字段、包含条件判断、或需要跨记录/跨时间计算的统计量，如 SUM,AVG,COUNTDISTINCT,MAX,MIN,PERCENTILE 等，或涉及多层分组、复杂过滤后的聚合，以及涉及去重计数（COUNTDISTINCT）、条件计数或多维度交叉计数的统计；

——应用数学公式或复杂业务逻辑计算生成新指标或字段：计算结果值在原始数据中不存在，且计算过程涉及 3 个及以上原始字段、包含条件判断分支、或需要跨记录/跨时间计算的业务逻辑；

——应用机器学习模型、统计模型生成预测值、评分或分类标签：模型的输入是原始数据或其特征，输出是模型推理产生的新信息（如预测用户是否购买、计算信用评分、识别图片内容）；

——对文本、图片、音频、视频等非结构化数据进行解析、特征提取、内容分析：如分词、实体识别、情感分析、主题建模、图像识别、语音转文本、特征向量化等将非结构化内容转化为结构化或可量化的信息；

——基于复杂条件进行的数据转换、清洗、标准化：标准化过程产生了新的分类或编码体系，或清洗/转换逻辑引入了新的业务含义或显著改变了数据分布。例如，将连续数值分箱为离散等级，将自由文本标签映射到标准标签体系；

——时间序列数据的重采样、复杂特征提取：如计算移动平均、趋势、波动性、傅里叶变换等指标，将原始时间点数据转化为时间段特征或频域特征；

——数据增强或数据合成：通过旋转、裁剪、SMOTE 等技术对原始数据进行变换或生成新样本，用于机器学习训练，如果这些操作改变了数据的统计分布或引入了新的特征。

7.1.3.3 评估操作对数据的影响

应评估识别出的关键操作对输入数据产生的改变程度。判断输出数据是输入数据的简单结果物(如选子集、格式转换)，还是通过专业知识加工、建模分析、关键信息提取等方式实现数据内容、形式、结构等实质改变。

7.1.4 判定准则

a) 若处理过程分析发现认定对象的生成依赖于输入数据，且过程中执行了 7.1.3.2 中所列的任何一类关键数据转换操作，则将认定对象初步判定为衍生数据；

b) 若处理过程仅涉及数据格式转换、编码转换、简单的字段更名、数据物理传输、数据复制、或基于简单条件（如 WHERE 子句）的数据筛选（仅生成输入数据的子集），未执行 7.1.3.2 中所列关键数据转换操作，则不应仅凭此认定为衍生数据。但应结合其他方法(如数据依赖关系分析、数据特征比对)进行综合判断。

7.1.5 判定结论

应结合判定过程，编制详细描述认定对象生成过程的数据处理过程分析报告，包含识别出的关键处理操作、对处理逻辑的评估，以及基于此分析得出的初步判定结论。

7.2 数据依赖关系分析法

7.2.1 适用范围与前提条件

本方法适用于能够追溯数据血缘信息或通过分析代码、配置识别数据依赖关系的场景，用于确定认定对象是否来源于其他数据。

前提条件：应当能够获取认定对象与其输入数据之间的数据血缘信息、ETL元数据、数据库依赖关系、API调用记录或可解析的数据流配置。

7.2.2 应获取到的关键素材

——数据血缘图；

——ETL 作业元数据；

——数据库表依赖关系图；

——API 调用日志；

——数据管道配置信息；

——数据流管理系统界面；

——描述数据流的文档。

7.2.3 处理过程

7.2.3.1 获取和构建数据血缘关系

利用现有数据治理工具、日志分析、代码解析或人工分析等方式，获取或构建认定对象的数据血缘信息，追溯源数据。

7.2.3.2 明确源数据

要求提供源数据，明确源数据组成，包括原始数据或其他具有使用权的数据。

7.2.3.3 追溯源数据

从认定对象开始，持续向上追溯，识别数据处理链条中的关键中间数据，并最终追溯到源数据。

7.2.3.4 分析依赖关系性质

应分析认定对象对输入数据的依赖关系性质，例如：

——内容依赖：认定对象的值来源于输入数据的计算或聚合；

——结构依赖：认定对象的结构由输入数据或处理逻辑决定；

——存在依赖：认定对象的存在性或记录数取决于输入数据。

7.2.4 判定准则

a) 若数据依赖关系分析证明认定对象在内容或结构上明确来源于(即通过处理过程从)一个或多个其他数据作为输入，且该生成过程涉及了 7.1.3.2 中所列的关键数据转换操作，则将认定对象初步判定为衍生数据；

b) 若认定对象仅是对源数据的物理复制、传输或简单的基于源数据的子集筛选(未伴随 7.1.3.2 中所列的关键转换)，则不应仅凭此判定为衍生数据；

c) 若无法追溯到明确的源数据，或者其唯一来源是自身(例如日志数据)，则不应判定为衍生数据(应视为原始数据)(如处理过程分析显示其是通过复杂计算或模型生成)。

7.2.5 判定结论

应形成认定对象的数据依赖关系图或清单，详细说明其直接和间接输入来源，以及基于此分析得出的初步判定结论。

7.3 数据特征比对法

7.3.1 适用范围与前提条件

本方法适用于能够获取认定对象、声明的和潜在的源数据样本，并通过比对数据特征来辅助判定。特别适用于处理过程和血缘信息不完全透明，但数据前后变化显著的场景。

前提条件：须能够获取认定对象及其声明的和潜在的源数据的样本；宜能够获取其数据字典或模式定义。

7.3.2 应获取到的相关素材

——认定对象数据集样本；

——声名的和潜在的数据集样本；

——数据字典；

——模式定义文件；

——统计分析工具；

——数据比对脚本或工具。

7.3.3 处理过程

7.3.3.1 确定比对维度

应根据数据类型（结构化、非结构化等)和预期的数据处理效果，选择合适的比对维度。

对于结构化数据，应比对字段数量、字段名称、数据类型、模式定义(如主外键关系、索引)、记录数量、数据粒度、各字段的统计摘要(如均值、中位数、方差、最大值、最小值、唯一值计数、缺失值率)、数据分布形态等。

对于非结构化数据，宜比对文件大小、文件类型、元数据、内容摘要、关键词分布、文本长度分布、图像分辨率、音频时长等。

7.3.3.2 执行数据特征比对

应使用合适的工具、编程脚本或统计分析方法，在选定的维度上对认定对象同声名的和潜在的源数据样本进行定量或定性比对。

7.3.3.3 分析并量化差异

详细分析比对结果，识别两数据集之间的差异，并对差异进行量化描述。

7.3.3.4 结合处理过程或依赖关系分析判断

将数据特征比对结果与已知的数据处理过程分析结果(参见7.1)或数据依赖关系分析结果(参见7.2)相结合。判断观察到的显著差异是否与已知的加工、分析、整合等处理活动一致。

7.3.4 判定准则

a) 若数据特征比对显示认定对象与源数据在以下任一维度存在显著差异，且该差异能够通 7.1 识别的关键处理操作 7.2 追溯到的复杂依赖关系得到合理解释，则应支持认定为衍生数据：
   1) 数据结构发生根本性改变(例如从多表扁平化为宽表，或从结构化到图结构)；
   2) 数据粒度发生根本性改变(例如从单笔交易记录聚合为用户月度消费汇总，或从传感器每秒读数聚合为分钟平均值)；
   3) 核心内容类型发生根本性改变(例如从原始数值到计算指标，从文本到情感标识，从图像像素到特征向量)；
   4) 关键统计特性(如均值、方差、分布形态)或数据分布发生显著变化，且非简单选或抽样所致。
b) 若比对发现仅存在格式、编码、字段名称等表层变化，或认定对象仅为源数据的简单子集(如基于单字段过滤)，且无 7.1 所述的关键处理操作支持，则不应仅凭此判定为衍生数据；
c) 显著差异的判定标准应根据具体的业务语境和数据类型进行定义，并在认定记录中予以说明；
d) 显著差异的量化判定应优先采用信息熵量化分析法（见 8.2）。

7.3.5 判定结论

应形成详细的数据特征比对报告，包含选定的比对维度、具体比对结果(量化或定性描述)、差异分析，以及结合其他方法分析得出的初步判定结论。

7.4 人工复核与确认法

7.4.1 适用范围与前提条件

本方法是所有自动化或半自动化认定方法的补充和保障。须用于对自动化判定结果不确定、存在争议，或涉及高度抽象、依赖复杂业务理解、涉及重要敏感数据的场景进行最终确认。

前提条件：须有具备相关专业知识和经验的人员参与；须有前期自动化或半自动化方法的分析结果和支持性证据。

7.4.2 应获取到的关键素材

——由方法 7.1、7.2、7.3 产出的完整分析报告与原始证据（如数据处理流程图、代码片段、血缘关系图、特征比对量化结果）；

——所有经识别的关键数据转换操作的详细描述及对应证据；

——认定对象、核心源数据及中间数据的样本；

——相关的业务规则文档、算法模型说明、数据字典及合规性要求文件；

——前期自动化判定过程中产生的不确定性或争议点记录。

7.4.3 处理过程

7.4.3.1 组织复核小组

组建专家复核小组，小组成员应至少涵盖数据治理、相关业务领域、技术实现及法律合规等领域。其中，业务领域与技术领域专家应具备至少三年相关经验，且全程独立、公正地参与。

7.4.3.2 审查判定依据

复核小组成员应当全面审查前期方法(参见7.1至7.3)得出的初步判定结论及其所依据的所有证据、分析报告和原始信息。

7.4.3.3 讨论与综合判断

复核小组应当对复杂或争议的数据处理场景和判定依据进行深入讨论和分析。应结合业务语境、技术实现细节、法律合规要求和对衍生数据定义的深刻理解，进行综合判断。

必要时，应要求数据处理人员、业务人员提供进一步解释、数据样本或系统演示。

7.4.3.4 形成最终结论

复核小组应当通过协商或必要的表决机制，形成一致的或按既定程序形成的最终认定结论。

该结论是本次认定的最终有效结果。

7.4.3.5 详细记录复核过程

应详细记录人工复核的过程，包括参与人员审查的依据、讨论的主要观点、存在的分歧意见及其理由、最终结论及其确定的理由。

7.4.4 判定准则

a) 若复核小组综合判断认为，前期证据充分、可靠，且一致支持认定对象是通过 7.1.3.2 所列的关键数据转换操作生成，并生成了新的信息内容、结构或含义，则应最终确认为衍生数据。

b) 若复核小组发现前期证据存在重大缺失、矛盾，或对关键操作的认定存在根本性分歧，且无法通过补充材料解决，则应作出“证据不足，无法认定”的结论，并指明所需的关键证据。

c) 若复核小组综合判断认为，前期认定的“关键操作”实质上未改变数据的本质信息内容（如仅为格式转换、简单映射或无业务含义的编码），或认定对象仅为源数据的子集，则应推翻前期初步结论，判定为非衍生数据。

d) 专家资质与利益回避是结论有效性的前提。若发现小组成员不符合资质要求或存在可能影响公正判断的利益冲突，其参与形成的结论应视为无效。

7.4.5 判定结论

应形成详细的人工复核与确认报告，作为认定的最终依据。该报告必须完整包含以下内容：复核小组的成员构成、资质说明及利益冲突声明；所审查的全部素材清单；对前期分析中每个关键争议点的审查意见与判断理由；复核过程中的主要分歧点及解决方式；明确的最终判定结论及详尽、可追溯的论证过程；以及全体参与复核人员的签字或等效确认记录

8 实践使用

8.1 基于规则或模型法

8.1.1 适用范围与前提条件

本方法适用于存在大量相似数据处理场景，可以通过提取数据元数据、处理元数据或数据特征，建立自动化或半自动化判定规则或模型进行高效判定的场景。

前提条件：必须能够获取用于规则匹配或模型输入的元数据或数据特征；必须已建立一套经过验证的判定规则集或训练好的判定模型。规则或模型的构建应基于前述方法的判定逻辑。

8.1.2 应获取到的素材

——预定义的业务规则集；

——判定逻辑模型；

——训练好的机器学习模型；

——数据元数据；

——处理过程元数据(如 ETL 作业类型、转换操作列表)；

——数据特征向量。

8.1.3 处理过程

8.1.3.1 获取判定特征

从认定对象及其生成过程的相关元数据(如源数据数量、处理作业类型、关键操作类型)或已提取的数据特征中，提取用于规则匹配或模型输入的特征。

8.1.3.2 应用判定规则或模型

将提取的特征输入到预定义的规则引擎或判定模型中。

规则引擎应执行匹配判断；模型应执行推理计算。

8.1.3.3 输出自动化判定结果

应记录规则匹配的结果(触发的规则)或模型输出的判定标签(如“衍生数据”“非衍生数据”)和置信度。

8.1.4 判定准则

a) 若规则匹配结果显示认定对象符合预定义的衍生数据判定规则(这些规则必须基于 7.1、7.2、7.3 判定标准的抽象和归纳)，则应将其判定为衍生数据；
b) 若判定模型输出的分类结果为衍生数据，且置信度高于预设阈值，则应支持将其判定为衍生数据；
c) 基于规则或模型的判定结果可作为初步结论，对于重要或敏感数据，或规则/模型输出置信度较低时，必须结合人工复核进行最终确认。

8.1.5 判定结论

应整合基于规则或模型自动化/半自动化得出的判定结论,以及触发的规则或模型的输出(如置信度、匹配规则详情)作为判定依据。

8.2 信息熵量化分析法

8.2.1 适用范围与前提条件

本方法是客观、优先的定量分析方法，适用于可获取认定对象及其输入数据样本的场景，作为数据特征比对法的核心量化工具和处理过程分析法的客观验证。

8.2.2 应获取到的素材

宜参照数据特征比对法的相关内容（7.3.2）。

8.2.3 处理过程

8.2.3.1 数据预处理与特征提取

根据数据类型（结构化、非结构化），对数据进行必要的离散化（如分箱）或特征提取（如词频统计、像素灰度直方图），使其适于概率分布计算。详细流程见附录F.2。

8.2.3.2 构建概率分布

基于预处理后的数据,分别计算原始数据集和衍生数据集在相应维度上的概率分布P_orig和P_der。

8.2.3.3 计算熵与相对熵

根据附录F.1中定义的公式，计算原始数据信息熵H_orig、衍生数据信息熵H_der、信息熵变化率ECR和相对熵D_KL(P_der || P_orig)。

8.2.3.4 阈值比对

将计算出的ECR和D_KL与附录F.3中针对不同场景推荐的阈值 δ 和 ε 进行比较。

8.2.4 判定准则

若相对熵D_KL超过预设阈值 ε，或信息熵变化率ECR超过预设阈值 δ，则应判定认定对象与源数据存在显著差异，强烈支持将其认定为衍生数据。

8.2.5 判定结论

判定结论可作为自动化认定的初步结论，并为人工复核提供客观的量化依据。

8.3 新兴技术与边界场景认定

对于AI生成内容（AIGC）、联邦学习、安全多方计算、差分隐私、数据增强/合成等新兴技术和边界场景，其认定应优先适用第7章的通用方法，并结合以下原则进行分析。

8.3.1 AI 生成内容（AIGC）

AIGC的生成过程通常涉及复杂的模型计算和对大量训练数据的学习。认定重点应在于识别其生成过程是否使输出结果的形态、结构、内容和含义与直接输入存在显著差异。

8.3.2 联邦学习/安全多方计算/差分隐私

这些技术旨在实现数据不出域情况下的协同计算或发布经过隐私处理的数据。

a) 最终计算结果/模型：无论过程多么注重隐私保护，最终汇聚或训练产生的模型参数、聚合报告、查询结果等，如果其生成过程涉及了 7.3.1.2 的关键转换操作（如多方数据聚合、模型训练/推理），且与任何单方原始数据存在显著差异，则应判定为衍生数据；

b) 中间交换信息：在计算过程中多方交换的加密数据、梯度、中间统计量等，如果它们主要用于后续计算流程，不作为独立资产提供，且其处理过程未达到生成最终结果的程度，则应参考中间数据的定义进行判断；

c) 判定重点：聚焦最终输出结果的生成过程是否符合 7.1.4，而非过程是否具备隐私保护特性。

8.3.3 数据增强/数据合成

专门用于机器学习训练的数据增强（如图片旋转、裁剪、亮度调整）或数据合成（如SMOTE生成合成样本），如果这些操作改变了数据的统计分布、引入了新的特征维度或生成了原始数据集中不存在的样本，则增强/合成后的数据应判定为衍生数据。如果仅是简单的复制或裁剪（未改变分布或引入新特征），则不应判定为衍生数据。

8.3.4 边界案例

对于难以明确判断的边界案例，必须依靠分级人工复核机制，由多领域专家综合权衡处理过程的实质性、结果数据的增值程度、业务场景和潜在风险，进行审慎判断。

9 认定流程

9.1 认定启动与计划

根据数据资产管理、安全合规、业务需求等触发条件，明确界定认定对象标识、认定范围和任务计划。可初步使用附录D的快速判定决策树进行方向性判断。

9.1.1 主要考虑因素

应将以下要素纳入认定计划：

——数据的重要性、敏感性级别；

——数据处理的复杂性；

——可获取的信息充分性；

——认定工作所需资源(时间、人力、工具)等。

9.1.2 认定计划

应形成明确界定的认定对象标识和认定任务计划。

9.2 认定相关信息的获取和梳理

应系统地收集与认定对象生成过程相关的所有必要信息。应包括认定对象的元数据、描述其生成逻辑的数据处理脚本或代码、系统配置信息、ETL作业定义、数据流图、数据血缘信息(如有)、相关技术文档、业务规则说明以及必要时对数据处理人员和业务专家的访谈记录。

9.2.1 主要考虑因素

宜在信息获取过程中，考虑以下因素：

——信息的完整性、准确性和可访问性；

——数据血缘工具的有效性；

——数据不可逆性；

——人工访谈的关键问题设计。

9.2.2 形成工作输出

应在本阶段结束后形成结构化或非结构化的信息集合，包括处理脚本、文档、访谈记录、初步梳理的数据流路径等。

9.3 核心判定分析

基于收集到的信息，识别关键数据转换操作和数据依赖关系，并可根据需要比对输入输出数据特征，应用规则或模型进行辅助分析，以形成对认定对象是否为衍生数据的初步判断。

9.3.1 主要考虑因素

宜在判定过程中，考虑以下因素：

——数据处理链条的复杂性；

——数据来源的多样性；

——数据类型的非结构化程度；

——现有工具和技术对处理过程和血缘分析的支持能力；

——所需比对的数据特征维度。

9.3.2 初步判定

应在本阶段后形成初步判定文档，包括但不限于数据处理过程分析报告、数据依赖关系图/清单、数据特征比对报告、基于规则/模型的判定结果和初步认定结论及其依据。

9.4 人工复核与最终确认

判定结果不确定、存在争议或涉及重要/敏感数据时，必须组织由具备相关专业知识的专家组成复核小组。

复核小组必须审阅分析报告、初步结论及其依据，进行讨论和综合判断。

必要时，应要求数据处理人员提供进一步解释或演示。

9.4.1 复核小组组建的要素

——复核小组成员的专业覆盖度(业务、技术、法律合规、数据治理等)；

——复核流程的规范性；

——分歧意见的协调机制。

9.4.2 建立分级复核机制

一般场景由3名专家复核；重要敏感数据（如涉及大量个人信息、国家重要数据、对业务决策有关键影响的数据）由至少5名专家组成复核小组。

9.4.3 复核结论

复核小组应给出复核会议纪要和形成的最终认定结论。

9.5 认定结果记录与归档

应以结构化、规范化的方式详细记录整个认定过程的关键信息和最终认定结论。

已完成的认定结果应妥善存储和管理。

9.6 认定结果应用与管理

认定结果应作为数据治理活动的基础。宜进行分类分级管理，并基于分级结果，将其应用于后续的数据资产目录更新、差异化安全策略制定、数据流通权限管理等活动中。

应建立认定结果的版本管理机制，定期或在认定对象的数据处理逻辑、源数据发生重大变化时对认定结果进行复核和更新。

9.7 争议处理

当认定结论存在争议时，应依照附录E的争议处理机制进行处理。

10 认定结果表述

衍生数据认定结果必须以清晰、规范、易于理解和追溯的方式进行记录和表达。

10.1 认定任务标识

应包含认定任务的唯一标识符。

10.2 认定对象标识

应明确指出被认定的具体数据集、数据表、文件或数据项的唯一或清晰标识符(如：数据资产名称、ID、存储路径、业务系统名称、表名等)。

10.3 认定结论

应给出明确的判断结果，例如：

——认定为衍生数据；

——认定为非衍生数据；

——尚待进一步认定。

10.4 依赖数据标识

如果认定对象被判定为衍生数据，必须明确源数据(包括具有合法使用权的原始数据或其他数据)的标识符或描述。

10.5 处理过程描述

应简要描述生成认定对象的关键数据处理过程和逻辑，例如：数据来源、主要转换/聚合/分析类型(如JOIN、GROUPBY、模型计算、文本分析)、使用的主要算法或模型名称。

宜引用相关的处理脚本、配置或文档的链接。

10.6 认定依据

应列出认定结论所依据的主要事实、证据来源和分析结果。

依据的引用应具体到文档名称、版本或链接。

若使用信息熵法，应记录D_KL和ECR的计算值及所用阈值。

10.7 采用的认定方法

应说明认定主要使用的方法/方法组合及原因。

10.8 认定执行者

应记录进行认定的主要人员或自动化系统的标识。

10.9 认定日期

应记录认定完成的日期。

10.10 认定版本与复核信息

宜使用版本号对认定结果进行管理。

若进行过人工复核，必须记录复核人员或复核小组的标识、复核日期。

附 录 A
（资料性）
通信行业衍生数据认定示例

A.1 场景概述

通信行业作为国家重要的基础设施，其业务运行和管理天然伴随着海量、高并发、多源异构的数据产生。这些数据包括用户通信记录(通话、短信)、上网行为日志(流量、时长、访问内容)、设备信令、网络性能指标、位置信息、客户资料等。通信行业对数据的使用需求多种多样，例如：网络优化与规划、用户行为分析、增值业务开发。风险控制、满足监管要求等。

然而，通信行业的数据处理过程也面临诸多挑战：数据来源复杂、处理链条长且隐蔽、涉及个人信息保护、数据量巨大且实时性高等。

在这样的背景下，准确识别和认定通信行业中产生的大量衍生数据，对于数据资产管理、合规经营、数据安全和价值挖掘至关重要。

A.2 示例选取

为深入阐述认定过程，选取两个通信行业的典型场景进行分析。

A.2.1 场景一：基于手机信令数据分析区域人流密度

原始数据：基站接收到的原始手机信令记录。

衍生数据：特定区域在不同时间段的人流密度统计报告。

A.2.2 场景二：基于用户上网行为日志生成用户兴趣画像

原始数据：经过初步处理和去标识化/匿名化的用户上网日志。

衍生数据：用户兴趣标签集合。

A.3 认定过程

A.3.1 启动认定

明确认定对象是区域人流密度报告数据集或用户兴趣标签数据集。

A.3.2 信息收集与梳理

收集生成目标数据集的ETL脚本、数据处理代码、数据流图、血缘信息、业务规则等。

A.3.3 核心判定分析

应用数据处理过程分析方分析脚本/代码，识别关键转换操作(如JOIN,GROUPBY,COUNTDISTINCT，模型应用)。

A.3.4 源数据追溯

应用数据依赖关系分析方法追溯数据来源至原始信令/上网日志。

A.3.5 判定活动

综合分析结果，应用数据特征比对方法、基于规则或模型的判定方法，并必须进行人工复核与确认以得出最终结论。例如，比对原始信令/日志与衍生数据的结构、粒度、内容差异，确认是否符合判定标准。

A.3.6 量化判断标准

A.3.6.1 处理过程复杂性

a) 关键操作计数：计算数据处理链中包含的前述所列出的关键操作类型数量（如关联、聚合、计算新指标、模型应用）

阈值：包含≥2 种关键操作类型，初步倾向于衍生数据。

f) 聚合粒度变化：原始数据通常是秒级/米级的个体事件，衍生数据是区域/时间段的汇总。量化粒度缩小倍数（原始时空单元数/衍生时空单元数）。

阈值：时空粒度聚合度>某个预设值（例如，原始数据每秒记录，衍生数据每小时区域汇总，聚合度极高），强烈倾向于衍生数据。

g) 计算指标复杂性：计算新指标涉及的原始字段数量、计算公式复杂性（简单求和 vs.复杂加权平均/指数计算）。

阈值：计算指标涉及≥3 个原始字段，或使用了非线性/跨字段计算，初步倾向于衍生数据。

h) 数据脱敏/匿名化方法：应用的脱敏或匿名化方法是否涉及复杂计算或显著改变数据形态（如 k-anonymity,差分隐私）。

阈值：应用 k-anonymity(k≥5)或差分隐私等方法，初步倾向于衍生数据（这些方法本身通常涉及复杂的数据变换）。

A.3.6.2 结果数据转变

a) 结构变化度量：字段数量变化率(|Output_Fields|-|Input_Fields|)/|Input_Fields|，模式复杂度变化（如从平面表到包含多级聚合指标的表）。

阈值：字段数量变化率>20%，或引入了至少一个计算指标，强烈倾向于衍生数据。

i) 数据粒度度量：结果数据不再是个体行为记录，而是群体统计值。比对原始数据记录数与结果数据记录数（通常结果记录数远小于原始记录数）。

阈值：结果记录数<原始记录数的 1%，强烈倾向于衍生数据（体现了高度聚合）。

j) 内容特征差异度量：结果数据的字段值是统计值或指标，而非原始的位置、时间、设备 ID。

阈值：结果数据的核心业务字段(>50%的非 ID 字段)是通过聚合或计算产生的新指标/统计值，强烈倾向于衍生数据。

A.3.7 判定检查表

检查项	描述	判定标准（量化）	初步结论权重
包含关键操作	是否包含聚合、计算新指标、模型应用等关键转换	≥2类关键操作：+3	强
聚合粒度	是否从个体细粒度数据聚合到区域/时间粒度统计	聚合度>100:1(粗略估算原始数据/衍生数据记录数比例）：+3	强
新指标计算	是否计算了原始数据中不存在的新指标	计算指标≥1个且涉及≥3个原始字段或公式复杂：+2	中
脱敏/匿名化	处理过程是否包含复杂脱敏或匿名化方法	是(应用k-anonymityk≥5或差分隐私等)：+1	弱
结构变化	结果数据与原始数据相比，字段数量、模式是否有显著变化	字段数量变化率>20%或引入新计算指标字段：+3	强
粒度变化	结果数据记录数是否远小于原始数据（体现聚合）	结果记录数<原始记录数1%：+3	强
内容转变	结果数据的核心业务字段是否为统计值/指标	>50%非ID字段为新统计值/指标：+3	强

将各检查项的权重相加。总权重≥5分，初步判定为衍生数据；总权重<3分，初步判定为非衍生数据；3≤总权重<5分，判定不确定，需加强人工复核。

A.3.8 人工复核要点

a) 数据脱敏有效性评估：即使已脱敏，复核小组需审视聚合粒度是否过细，是否仍存在“逆向别”出个体的潜在风险。确认聚合结果是否真正消除了个体轨迹信息。特别注意 k-anonymity 等方法的参数选择（如 k 值）是否符合既定标准（如 k≥5），以及这些方法是否真正改变了数据的底层结构或可识别性。

k) 空间聚合算法的实质性：审视空间聚合算法的复杂度和创新性，区分简单的区域筛选（非衍生）与复杂的空间插值、热力图生成或多区域关联分析（倾向衍生）。

l) 时间窗口选择的影响：评估时间窗口选择是否引入了新的时间序列特征或统计意义（如高峰低谷计算），而非简单的按时间段分割。

m) 指标计算的业务含义：理解计算指标（如“人流密度指数”）的业务意义，判断它是否代表了对原始数据的新洞察或新维度，而非原始字段的简单代数运算。

n） 数据血缘的完整性：确认收集到的数据血缘信息是否完整，是否存在未被文档或脚本覆盖的关键处理环节。

o） 规则或模型应用的审视：如果过程中使用了基于规则的过滤或简单的统计模型，复核这些规则或模型的设定是否引入了实质性的数据转换或新的业务含义。

A.3.9 结果记录与输出

按照本文件第10章要求，结构化记录认定过程和结论。

A.4 案例分析与认定结果表述

A.4.1 案例一：区域人流密度报告认定

a） 原始数据：手机信令采集系统导出的去标识化信令明细数据。

p） 关键数据处理步骤：对原始信令进行清洗、去重，与基站地理信息关联，按区域和时间粒度进行聚合统计(COUNTDISTINCT)，计算人流密度指数。

q） 最终生成的衍生数据：数据仓库中包含区域、时间段、活跃终端数、人流密度指数等字段的统计报告表。

r） 认定为衍生数据的关键理由和依据：数据经过了复杂的关联、聚合统计和计算新指标的操作，从细粒度的事件记录转化为区域层面的统计汇总数据，结构、粒度、内容与原始数据显著不同，符合本文件 7.1、7.2、7.3 的判定标准。

A.4.2 案例二：用户兴趣标签认定

a） 原始数据：用户上网行为日志平台导出的去标识化/匿名化日志数据。

s） 关键数据处理步骤：对上网日志进行清洗、标准化，按用户分组统计行为特征(如访问时长、流量)，将特征输入机器学习模型，模型输出用户兴趣标签或评分。

t） 最终生成的衍生数据：用户画像库中用户兴趣标签字段或兴趣评分字段。

u） 认定为衍生数据的关键理由和依据：数据经过了复杂的特征提取和机器学习模型处理，将离散、细粒度的日志记录转化为用户层面的高维特征和抽象标签，结构和内容与原始日志完全不同，符合本文件 7.1、7.2、7.3 的判定标准。

附 录 B
（资料性）
电商平台衍生数据认定示例

B.1 场景概述

电商平台是典型的以数据驱动业务的场景。在用户、商品、交易、营销等各个环节都会产生海量数据。这些数据类型多样，主要包括：用户数据、行为数据、交易数据、商品数据、评价数据、营销活动数据、日志数据等。

电商平台的数据处理具有高并发与实时性、多源异构、复杂关联与聚合、模型驱动应用等特点。面临的挑战包括数据处理链条长、处理逻辑复杂且动态、技术实现多样性、区分中间数据与衍生数据等。

在电商平台中，准确认定衍生数据对于数据资产管理、数据安全与合规、数据流通与交易、业务决策与优化具有极其重要的意义。

B.2 示例选取

为深入阐述认定过程，选取两个电商平台的典型场景进行分析。

B.2.1 场景一：个性化商品推荐列表

原始数据：用户历史行为数据、商品基础信息。

衍生数据：为特定用户实时或周期性生成的个性化商品推荐列表。

B.2.2 场景二：用户消费能力分级

原始数据：用户基本信息、历史交易数据、支付记录、会员等级信息等。

衍生数据：针对用户生成的消费能力等级或潜在消费能力评分。

B.3 认定过程

B.3.1 启动认定

明确认定对象是推荐列表数据或消费能力分级/评分数据。

B.3.2 信息收集与梳理

收集推荐/评分算法代码、特征工程脚本、ETL配置、数据流图、业务规则等。

B.3.3 核心判定分析

应用数据处理过程分析方法分析代码/配置，识别关键操作(如特征提取、模型预测、规则匹配、聚合统计、多源关联)。

B.3.4 源数据追溯

应用数据依赖关系分析方法追溯数据来源至用户行为日志、交易数据、商品数据等原始数据。

B.3.5 判定活动

综合分析结果，应用数据特征比对方法、基于规则或模型的判定方法并必须进行人工复核与确认以得出最终结论。例如，比对原始日志/交易数据与衍生数据的结构、内容、粒度差异，确认是否符合本文件7.1.4至7.4.4的判定标准。对于涉及个人敏感信息的场景，人工复核尤为重要。

B.3.6 量化判断标准

B.3.6.1 处理过程复杂性

a) 算法类型评分：

1) 基于简单规则/阈值过滤排序：1 分

2) 基于协同过滤（用户/物品）、内容推荐、基本因子分解机（MF）：3 分

3) 基于复杂机器学习模型（如 GBDT,SVM）、浅层神经网络：5 分

4) 基于深度学习模型（如 DNN,RNN,Transformer）、复杂图神经网络：8 分

5) 多阶段召回+复杂排序模型：10 分

6) 阈值：算法类型评分≥5 分，强烈倾向于衍生数据。

v) 特征工程复杂性：用于模型的输入特征数量；特征是否经过复杂组合、交叉或嵌入。

阈值：输入特征数量≥20 个或包含复杂组合/嵌入特征：+3 分。

w) 数据融合度：生成推荐列表是否整合了用户行为、商品属性、用户画像等≥3 种不同维度的源数据。

阈值：数据融合度≥3 个维度：+2 分。

B.3.6.2 结果数据转变

a) 结构变化：结果数据不再是原始行为记录或商品属性，而是针对特定用户的有序商品列表。

度量：结果数据结构是定长的推荐列表而非变长的事件记录或固定结构的商品信息。结构改变度高：+3 分。

x) 内容转变（个性化度）：推荐列表是基于特定用户的历史行为和偏好生成的，与其他用户的列表或全局热门列表存在差异。

度量（需抽样评估或系统计算）：

y) 与全局热门商品列表的重叠率：重叠率<50%：+2 分；重叠率<20%：+4 分。

与随机抽取用户推荐列表的平均重叠率（如果系统支持计算）：平均重叠率<30%：+3 分；平均重叠率<10%：+5 分。如系统不支持该计算，可采用抽样对比方式，随机抽取至少 10 个用户的推荐列表与认定对象（某个用户的推荐列表）进行人工比对或计算简单重叠率进行辅助判断。

阈值：个性化度量得分≥3 分，强烈倾向于衍生数据。

z) 新信息/含义：推荐列表本身代表了一种新的“用户-商品关联”预测，原始行为数据中没有这种显式的预测关联。

度量：结果数据是否包含推荐分数、推荐理由等原始数据不包含的“预测性”信息。包含此类信息：+3 分。

B.3.7 判定检查表

检查项	描述	判定标准（量化）	初步结论权重
算法类型	使用的推荐算法复杂程度评分	评分≥5：+5	强
特征工程	输入特征数量及复杂性	特征数量≥20或含复杂特征：+3	中
数据融合	融合源数据维度数	≥3种维度：+2	中
结构变化	结果是针对特定用户的有序推荐列表	结构改变度高：+3	强
内容转变(个性化)	推荐列表与热门/随机列表差异度	个性化度量得分≥3：+3	强
新信息/含义	结果包含预测性信息(分数/理由)	包含：+3	中

将各检查项的权重相加。总权重≥8分，初步判定为衍生数据；总权重<5分，初步判定为非衍生数据；5≤总权重<8分，判定不确定，需加强人工复核。

B.3.8 人工复核要点

a) 算法原理实质性：审视推荐算法的核心原理，确认其是否通过对原始数据的深度分析、模式挖掘、预测建模等生成新的关联或预测结果，而非简单的基于原始行为的规则匹配（如“买了 A 就推荐 B”）。这与前述“模型计算”或“复杂业务逻辑”相对应。

aa) 特征输入的完整性与关联性：确认用于模型的输入特征是否完整反映了用户行为和商品属性，以及这些特征与推荐结果之间是否存在合理的业务逻辑关联（防止“伪造”的复杂性）。

ab) 模型“黑箱”审视：对于复杂的机器学习模型，复核小组需了解模型的输入、输出和核心逻辑（即使内部细节不完全透明），判断其是否具备从原始数据中学习并生成新预测信息的能力。

ac) 推荐结果的业务有效性：结合业务场景，评估推荐列表是否真正体现了对用户偏好的“理解”和预测，是否能通过解释推荐理由（如果提供）回溯到用户行为和商品特征（逻辑上的可追溯性）。

ad) 实时性与计算开销：实时推荐系统通常涉及复杂的实时特征计算和模型推理，这本身就体现了高强度的数据处理，是判断衍生性的重要辅助依据。

ae) 敏感信息的使用：确认特征工程和模型训练过程中是否使用了用户的敏感信息（如地理位置、健康偏好），这会提升数据处理的敏感度，要求更审慎的复核。

B.3.9 结果记录与输出

按照本文件第10章的要求，结构化记录认定过程和结论。

B.4 案例分析与认定结果表述

B.4.1 案例一：个性化商品推荐列表认定

a) 原始数据：用户行为日志数据、商品基础信息。

af) 关键数据处理步骤：提取用户行为特征，结合商品特征，应用召回算法和排序模型，生成个性化推荐列表。

ag) 最终生成的衍生数据：为特定用户生成的有序商品列表。

ah) 认定为衍生数据的关键理由和依据：数据经过复杂的特征提取、多算法召回、模型打分排序等处理，从分散的行为事件和商品信息中生成了为特定用户定制的新数据，结构、内容、粒度与原始数据存在根本性差异符合本文件 7.1、7.2、7.3 的判定标准。

B.4.2 案例二：用户消费能力分级认定

a) 原始数据：用户注册信息、历史订单记录、支付记录、会员等级信息。

ai) 关键数据处理步骤：整合多源数据，计算用户消费特征(总消费、订单频率等)，应用分级规则或评分模型生成消费能力等级或评分。

aj) 最终生成的衍生数据：用户画像库中用户的消费能力等级标签或评分字段。

ak) 认定为衍生数据的关键理由和依据：数据经过跨系统整合、多维度特征计算、聚合统计和规则/模型判断，从离散、细粒度的交易和基础数据中提炼出用户层面的综合经济评价指标，结构、内容、粒度与原始数据存在显著差异，符合本文件 7.1、7.2、7.3 的判定标准。人工复核确保了对涉及个人信息的敏感数据的审慎判定。

附 录 C
（资料性）
多源公开数据整合场景衍生数据认定示例

C.1 场景概述

多源公开数据整合场景是指从各种公开渠道(如政府部门网站、行业协会、上市公司公告、新闻媒体、科研机构发布的数据、公开API等)收集数据，经过清洗、标准化、关联、聚合、分析等过程，形成具有新价值的数据集或洞察报告。这类数据通常用于市场研究、竞争分析、政策评估、风险监控、社会趋势分析等领域。

该场景的数据处理具有来源极其分散多样、数据质量参差不齐、整合处理复杂、价值提炼依赖深入分析等显著特点。面临的挑战包括数据处理流程非标准化、处理逻辑隐蔽、原始数据易变性、区分简单汇聚与实质性衍生等。

在这样的场景下,准确识别和认定整合分析后产生的数据是否为衍生数据,对于明确数据资产边界、评估分析产出的价值、进行知识产权管理和合规使用具有重要意义。

C.2 示例选取

为深入阐述认定过程，选取两个多源公开数据整合的典型场景进行分析。

C.2.1 场景一：行业市场规模与竞争格局分析数据表

原始数据：来源于政府统计公报、上市公司年报、行业协会报告等。

衍生数据：经过整合、计算、标准化后形成的结构化数据表，包含各主要企业在特定年份在不同细分市场的关键指标。

C.2.2 场景二：特定主题公共舆情分析报告

原始数据：来源于新闻网站报道、社交媒体公开评论、行业论坛帖子等关于特定主题的文本数据。

衍生数据：经过文本清洗、分词、主题建模、情感分析等处理后形成的结构化数据，如主题热度趋势、情感指数等。

C.3 认定过程

C.3.1 启动认定

明确认定对象是市场分析数据表或舆情分析报告数据。

C.3.2 信息收集与梳理

收集原始公开数据来源(URL、文件清单)、采集解析脚本、清洗整合脚本、分析模型代码、计算逻辑文档等。

C.3.3 核心判定分析

应用数据处理过程分析方分析脚本/代码，识别关键转换操作(如抽取解析、清洗标准化、实体对齐、多源整合、计算指标、文本分析)。

C.3.4 源数据追溯

应用数据依赖关系分析方法追溯数据来源至原始公开文件、网页、API等。

C.3.5 判定活动

综合分析结果，应用数据特征比对方法、基于规则或模型的判定方法，并必须进行人工复核与确认以得出最终结论。例如，比对原始数据与衍生数据的格式、结构、内容、粒度差异，确认是否符合本文件7.1.4至7.4.4的判定标准。人工复核在处理公开数据质量不确定和复杂逻辑时尤为重要。

C.3.6 量化判断标准

C.3.6.1 源数据复杂性

a) 源数据数量：集成独立源数据的数量。

阈值：集成≥3 个独立源数据：+2 分；集成≥5 个独立源数据：+4 分。

a1) 源数据类型多样性：集成的原始源数据是否包含结构化、半结构化、非结构化等≥2 种不同类型。

阈值：集成≥2 种类型：+2 分。

C.3.6.2 处理过程复杂性

a) 清洗标准化复杂性：是否涉及跨格式（如从 PDF/HTML 抽取表格）、复杂的模式转换、非简单查找的标准化映射。

阈值：涉及跨格式抽取或复杂模式转换或非简单映射标准化：+2 分。

am) 实体对齐/融合复杂性：是否涉及模糊匹配、人工校对规则、复杂实体解析算法。

阈值：涉及模糊匹配或复杂解析算法：+3 分。

an) 整合/计算逻辑复杂性：是否涉及多表 JOIN(≥3 表)、跨源聚合计算、复杂新指标计算公式、文本分析模型应用。

阈值：涉及多表 JOIN(≥3 表)或跨源聚合或复杂新指标计算或文本分析模型应用：+4 分。

C.3.6.3 结果数据转变

a) 结构统一性：结果数据是否将多种异构源统一为规范的结构化格式。

度量：结果结构化程度显著高于任一原始源：+3 分。

ao）信息密度/洞察力：结果数据是否包含了在任一原始源中无法直接获取的、通过整合计算得出的新信息或聚合视图。

度量：结果数据中≥30%的核心业务字段是通过整合/计算产生的复合指标或聚合值：+4 分。

ap）数据粒度：结果数据是否从文档、文本等粒度聚合到实体（公司、事件、主题）或主题层面的统计/分析结果。

度量：数据粒度发生显著聚合：+3 分。

C.3.7 判定检查表

检查项	描述	判定标准（量化）	初步结论权重
独立源数量	集成的独立源数据个数	≥3个：+2;≥5个：+4	中
源类型多样性	集成源包含的类型(结构化/非结构化)	≥2种类型：+2	中
清洗标准化	处理复杂性(跨格式/复杂映射)	是：+2	中
实体对齐/融合	涉及模糊匹配/复杂算法	是：+3	强
整合/计算逻辑	涉及多JOIN/跨源聚合/复杂计算/文本模型	是：+4	强
结构统一性	结果是否高度结构化	是：+3	中
信息密度/洞察	新生成的复合指标/聚合值占比	≥30%核心字段是新生成：+4	强
粒度变化	结果粒度是否显著聚合	是：+3	中

将各检查项的权重相加。总权重≥10分，初步判定为衍生数据；总权重<6分，初步判定为非衍生数据（可能仅是简单汇聚）；6≤总权重<10分，判定不确定，需加强人工复核。

C.3.8 人工复核要点

a） 源数据独立性与可信度：确认所有声称的“独立源数据”确实是相互独立的公开数据发布渠道，而非同一数据的不同副本或格式。评估原始源数据的官方性、权威性和更新频率，这会影响整合结果的价值。

aq）整合逻辑的合理性与有效性：审查实体对齐规则是否准确、数据融合策略是否合理，确保整合过程没有引入错误或丢失关键信息。特别是模糊匹配等复杂规则，需要人工抽样验证效果。

ar）新计算指标的业务意义与算法：深入理解所有新计算指标的业务含义和计算方法，判断它们是否真正代表了对原始数据的有价值的提炼或综合分析结果，而非简单的指标组合。对于文本分析结果（如情感得分、主题分类），需要人工抽样检查模型的准确性。

as) 结果数据的增值体现在何处：复核小组需评估整合后的数据集是否回答了仅凭任一原始源数据无法回答的问题，是否为用户提供了新的分析视角或决策支持。这是判断“增值程度”和“实质性改变”的关键。

at) 潜在的数据质量问题：多源整合极易引入数据不一致、冲突、缺失等问题。复核小组需了解数据清洗和质量控制措施，评估结果数据的可靠性。低质量的整合结果可能不应被视为高价值的衍生数据。

au) 简单汇聚与深度整合的界定：人工复核是区分简单汇聚（如将多份 Excel 表用 UNIONALL 合并）与深度整合（如通过实体对齐、复杂 JOIN、跨源计算生成新的关系型视图）的核心环节。重点关注处理过程中是否有非平凡的逻辑转换和信息提炼，是否符合前述关键操作的定义。

C.3.9 结果记录与输出

按照本文件第10章要求，结构化记录认定过程和结论。

C.4 案例分析与认定结果表述

C.4.1 案例一：行业市场规模与竞争格局分析数据表认定

a) 原始数据：国家统计局公报、上市公司年报、行业协会报告(PDF、Excel、网页等)。

av) 关键数据处理步骤：抽取、解析、清洗、标准化、实体对齐多源数据，按维度整合关联，计算市场份额、增长率等新指标，生成结构化分析表。

aw) 最终生成的衍生数据：数据仓库中的行业市场分析表，包含年份、公司、市场份额、增长率等计算指标。

ax) 认定为衍生数据的关键理由和依据：数据经过复杂的跨格式抽取、多源整合、清洗标准化和指标计算，从分散异构的原始公开数据中生成了结构化、标准化且包含计算指标的新数据，结构、内容、粒度与原始数据存在显著差异，符合本文件 7.1、7.2、7.3 的判定标准。

C.4.2 案例二：特定主题公共舆情分析报告数据认定

a) 原始数据：新闻网站、社交媒体、论坛等关于特定主题的原始文本。

ay) 关键数据处理步骤：采集、清洗、预处理文本，应用 NLP 模型(情感分析、主题模型)，提取特征，结构化结果并与元数据关联，按维度聚合统计，生成舆情分析报告数据。

az) 最终生成的衍生数据：分析平台中的舆情分析报告数据表，包含统计时间段、主题标签、平均情感得分等指标。

ba) 认定为衍生数据的关键理由和依据：数据经过复杂的文本处理和 NLP 模型应用，将海量非结构化文本转化为定量化、结构化的分析指标，结构、内容、粒度与原始文本存在根本性差异，符合本文件 7.1、7.2、7.3 的判定标准。

附 录 D
（资料性）
快速判定决策树

D.1 本决策树提供一个简化的判断流程，用于快速对数据进行初步归类。复杂或难以判断的情况仍需遵循完整的认定流程。

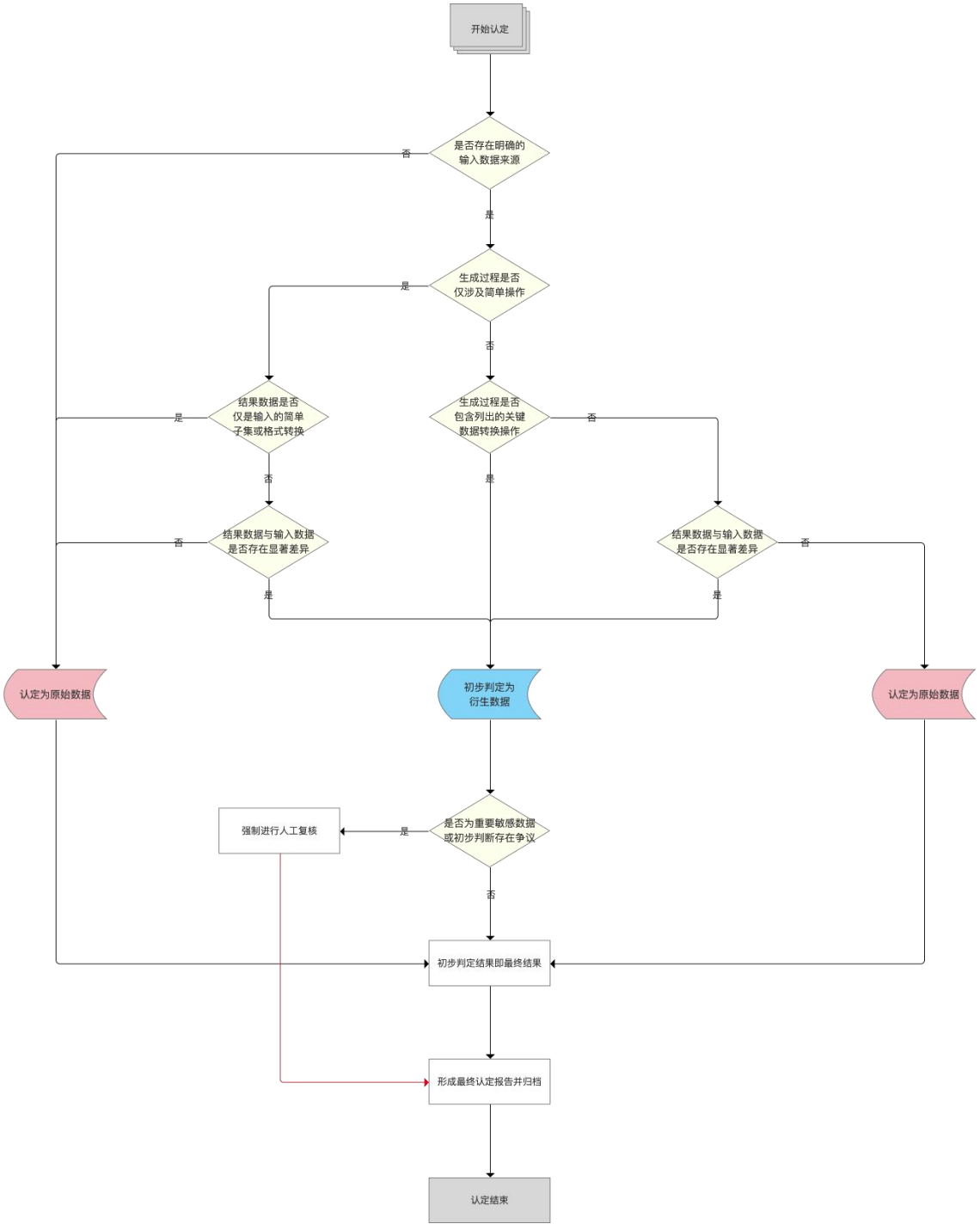


图 D.1　快速判定决策树示意图

附 录 E
（资料性）
争议处理机制

E.1 当衍生数据认定结论存在争议时，应启动本争议处理机制。

E.2 争议提出与受理

a) 对认定结论有异议的方（如数据所有者、数据使用者、数据处理者、合规部门等）应在收到认定报告后的规定时间内（例如，5 个工作日内）向指定的数据治理或争议处理机构（例如，数据合规委员会、法务部）提交正式的《衍生数据认定争议申请书》。

bb) 申请书应详细说明争议的事项、理由及支持性证据（如不同的处理过程描述、对标准的理解差异、新的信息等）。

bc) 争议处理机构负责对申请进行形式审查，确保申请材料的完整性。符合要求的应予受理并通知相关方。

E.3 争议复核与调解

a) 争议处理机构应组织原认定复核小组之外的、具备更高层级或更广泛领域专业知识的专家组成争议复核小组（人数可根据争议复杂性和数据重要性确定，不宜少于 3 人）。

bd) 争议复核小组应全面审阅原始认定报告、争议申请书及双方提供的所有相关材料。

be) 争议复核小组可以要求相关方进行说明、提供补充证据，并可以组织听证会或进行现场核查。

bf) 争议复核小组应尝试进行调解，促使争议各方达成一致。

E.4 专家仲裁与裁决

a) 如调解不成，争议复核小组应基于事实、本标准条文、相关法律法规和行业惯例，形成专家仲裁意见。

bg) 专家仲裁意见应详细说明采纳或驳回争议理由的原因，以及对原认定结论的维持或修改建议。

bh) 专家仲裁意见提交给争议处理机构决策层（如数据合规委员会主任、高级管理层）进行最终裁决。

bi) 最终裁决结果应以书面形式通知争议各方，并说明裁决理由。

E.5 结果执行与记录

a) 争议各方应执行最终裁决结果。

bj) 争议处理过程中的所有关键文件（申请书、复核意见、裁决书等）应完整记录和归档，作为该数据认定历史的一部分。

bk）如最终裁决修改了原认定结论，应更新该数据的正式认定记录和报告。

附 录 F
（资料性）
信息熵量化分析与应用指南

F.1 核心概念与数学定义

F.1.1 信息熵

对于一个离散型随机变量 X，其可能取值为 $\{x_1, x_2, \ldots, x_n\}$，每个取值的概率为 $p(x_i)$，其信息熵 H(X) 定义为：

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \quad \text{(F.1)}$$

单位为比特（bit）。

F.1.2 相对熵

用于衡量衍生数据的概率分布 P_der 相对原始数据概率分布 P_orig 的差异，定义为：

$$D_{KL}(P_{der}|P_{orig}) = \sum_i P_{der}(x_i)\log_2 \frac{P_{der}(x_i)}{P_{orig}(x_i)} \quad \text{(F.2)}$$

D_KL 值越大，表示两个分布差异越大。

F.1.3 信息熵变化率

用于衡量数据不确定性的相对变化幅度，定义为：

$$ECR = \frac{|H_{der} - H_{orig}|}{H_{orig}} \quad \text{(F.3)}$$

其中 H_orig 和 H_der 分别为原始数据和衍生数据的信息熵。

F.2 熵计算流程与方法

F.2.1 自动化认定流程

以下流程描述了如何自动化执行信息熵量化分析：

a) 输入与识别：输入原始数据和衍生数据样本，识别数据类型（结构化、非结构化）。

b1) 预处理：根据数据类型，执行 F.2.2 中相应的预处理操作。

bm) 概率分布计算：基于预处理后的数据，计算 P_orig 和 P_der。

bn) 核心指标计算：计算 H_orig, H_der, ECR 和 D_KL。

bo) 阈值比对：将计算结果与 F.3 中的推荐阈值进行比较。

bp) 输出报告：生成包含量化指标和初步结论的认定报告。

F.2.2 不同数据类型的熵计算方法

表 F.1 不同数据类型的熵计算方法

数据类型	描述	熵计算方法
结构化 - 离散/分类型	字段值为有限、可枚举的类别。	1. 统计每个类别的出现频率，得到概率分布 $p(x_i)$。 2. 直接应用香农熵公式计算。
结构化 - 连续/数值型	字段值为连续的数值。	1. 数据离散化：采用等宽分箱、等频分箱或聚类算法（如K-Means）将连续值映射到有限的“桶”中。 2. 对分箱后的离散数据计算熵。
非结构化 - 文本	自由格式的文本数据。	1. 特征提取：计算词频（TF）、n-gram频率或通过主题模型（如LDA）提取主题分布。 2. 基于提取的特征频率构建概率分布，计算熵。
非结构化 - 图像	像素数据。	1. 特征提取：计算图像的像素灰度直方图。 2. 将直方图归一化为概率分布，计算熵。

F.3 推荐阈值表

阈值 ε (KL散度) 和 δ (ECR) 的设定应根据具体业务场景、数据敏感度和风险容忍度进行调整。下表提供了一组示例性推荐值，用户应在实践中进行校准。

表 F.2 推荐阈值表

场景/数据敏感度	数据处理目标	相对熵阈值 ε	ECR阈值 δ	判定说明
低敏感度（如：常规统计报表）	聚合、汇总、简单清洗	> 0.5	> 30%	指标超过任一阈值，即认为存在显著差异。
中敏感度（如：用户画像标签、市场分析）	特征工程、规则引擎、多源融合	> 0.2	> 20%	阈值更敏感，较小的分布变化也可能被视为衍生。
高敏感度（如：金融风控评分、医疗诊断模型）	复杂机器学习模型、AIGC	> 0.1	> 10%	极低的阈值，因为微小的变化也可能代表信息内涵的巨大改变。
隐私计算（如：差分隐私发布）	隐私保护下的数据发布	> 1.0	> 50%	KL散度或ECR需足够大，以证明与原始分布存在巨大差异，确保隐私安全。

F.4 新兴技术场景应用指南

F.4.1 联邦学习

a) 认定逻辑。认定对象是最终的全局模型。其价值在于整合了多方信息，降低了对未知数据的预测不确定性（熵）。

bq）操作建议。使用标准测试集，比较全局模型预测分布的熵 H_global 与各局部模型预测分布的加权平均熵 H_local_avg。若 H_global 显著低于 H_local_avg，则证明发生了实质性改变，应判定全局模型为衍生数据。

参 考 文 献

[1] XXXXXXX

[2] XXXXXXX