ICS 35.240
L 72

DB37

山　　东　　省　　地　　方　　标　　准

DB 37/T 3523.2—2019

公共数据开放　第2部分：数据脱敏指南

Public data openness—Part 2：Data desensitization guidelines

2019－03－21 发布　　2019－04－21 实施

山东省市场监督管理局　　发布

目　　次

前　言

DB37/T 3523《公共数据开放》分为如下部分：

——第1部分：基本要求；

——第2部分：数据脱敏指南；

——第3部分：开放评价指标体系；

——第4部分：……

本部分为DB37/T 3523的第2部分。

本部分按GB/T 1.1—2009给出的规则起草。

本部分由山东省大数据局提出、归口并监督实施。

本部分起草单位：山东省大数据局、山东省公安厅、山东省计算中心（国家超级计算济南中心）、山东省大数据中心、山东省标准化研究院。

本部分主要起草人：柯林森、赵一新、李明、闫雷、赵硕、史丛丛、王洪儒、张媛、逄锦山、綦琳、陈洪波、李学民、刘晓飞、李刚、周鸣乐。

公共数据开放　第2部分：数据脱敏指南

1　范围

本部分提供了公共数据开放中数据脱敏的指导和建议，并给出了基本原则、脱敏规划、脱敏流程等方面需考虑的要点信息。

本部分适用于山东省公共数据开放的数据脱敏工作。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是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 35273　信息安全技术　个人信息安全规范

3　术语和定义

下列术语和定义适用于本文件。

3.1

数据脱敏　data desensitization

按照一定规则对原始数据进行处理，达到屏蔽敏感信息的一种数据保护方法。

3.2

个人敏感信息　personal sensitive information

一旦泄露、非法提供或滥用可能危害人身和财产安全，极易导致个人名誉、身心健康受到损害或歧视性待遇等的个人信息。

注：个人信息包括身份证件号码、个人生物识别信息、银行账号、通信记录和内容、财产信息、征信信息、行踪轨迹、住宿信息、健康生理信息、交易信息、14周岁以下（含）儿童的个人信息等。

[GB/T 35273，定义3.2]

4　基本原则

4.1　有效

数据脱敏宜确保脱敏工作的有效性，去除数据中的敏感信息，保证数据安全，经数据脱敏处理后，原始信息中包含的敏感信息已被移除，无法通过处理后的数据得到敏感信息，并防止使用非敏感数据推断、重建敏感原始数据。

4.2　真实

数据脱敏宜确保脱敏工作的真实性，脱敏后的数据应尽可能真实地体现原始数据的特征，且应尽可能多保留原始数据中的有意义信息。在开展数据脱敏工作时，一般情况下宜注意以下方面：

a) 保持原数据的格式；
b) 保持原数据的类型；
c) 保持原数据间的依存关系；
d) 保持语义完整性；
e) 保持引用完整性；
f) 保持数据的统计、聚合等特征；
g) 保持频率分布；
h) 保持唯一性。

4.3 高效

数据脱敏宜确保脱敏工作的高效性，宜通过程序自动化实现，并可重复执行。在不影响有效性的前提下，需注意平衡脱敏的力度与所花费的代价，将数据脱敏的工作控制在一定的时间和经济成本内。

4.4 稳定

数据脱敏宜确保脱敏工作的稳定性，需保证对相同的原始数据，在各输入条件一致的前提下，无论脱敏多少次，其最终结果相同。

4.5 可配置

数据脱敏宜确保脱敏工作的可配置性，按照输入条件不同生成不同的脱敏结果，从而可按数据使用场景等因素为不同的最终用户提供不同的脱敏数据。

5 脱敏规划

宜对数据脱敏工作进行总体规划，制定完备的数据脱敏工作方案，并对可能接触到脱敏数据的相关方进行数据脱敏规程的培训，并定期评估和维护数据脱敏规程内容。在制定数据脱敏工作方案时，宜考虑以下因素：

a) 明确敏感数据管理部门，以及其安全责任和义务；
b) 建立敏感数据的分类分级、脱敏工具运维管理等制度，并定期维护更新；
c) 建立数据安全管控机制，如代码安全、审计安全、安全管理等；
d) 定期对数据脱敏工作的相关方开展培训工作；
e) 制定完备的敏感数据使用审批流程，确保敏感数据的使用安全合规；
f) 明确数据脱敏流程，包括发现敏感数据、标识敏感数据、确定脱敏方法等。

6 脱敏流程

6.1 识别敏感数据

宜完整地梳理数据中包含的信息，明确其中敏感信息，识别敏感数据包括但不限于：

a) 明确数据脱敏工作范围；
b) 对工作范围内数据进行梳理和分类；
c) 建立敏感数据位置和关系库，以保存敏感数据的位置，以及敏感数据与原数据之间的关联关系；

d) 根据业务需要选择人工或自动等识别方式,并考虑识别方式与主流数据库系统、数据仓库系统、文件系统、云计算环境下新型存储系统等的适用性；
e) 选择数据发现工具，并考虑其扩展性，可根据业务需要自定义敏感数据的发现逻辑；
f) 明确敏感信息的字段名称、字段类型、字段长度、赋值规范等内容；
g) 利用反关联方法，查找可能由某些非敏感字段推断出另一敏感字段的映射，并对这些非敏感字段进行识别，例如：由出生日期可以推断出身份证号码的场景，需对出生日期进行识别。

6.2 标识敏感数据

识别出敏感数据后，宜尽早对敏感数据的格式、位置等信息进行标识，标识方法的选择宜考虑以下因素：

a) 敏感数据标识信息能够随敏感数据一起流动；
b) 敏感数据标识信息不易被恶意攻击者删除和篡改；
c) 需考虑便捷性和安全性，使标识后的数据容易被识别；
d) 需支持不同数据类型（如静态数据和动态数据）的敏感标识；
e) 对所有可能生成敏感数据的非敏感字段同样进行标识，例如：在病人诊治记录中为隐藏姓名与病情的对应关系，将“姓名”作为敏感字段进行变换，但是如果能够凭借某“住址”的唯一性导出“姓名”，则需要将“住址”进行标识并脱敏。

6.3 确定脱敏场景

在标识敏感数据基础上，确定脱敏场景，脱敏场景包括但不限于：

a) 静态脱敏：对原始数据进行一次脱敏后，脱敏后的结果数据可以多次使用；
b) 动态脱敏：针对不同用户需求，对数据进行屏蔽处理的数据脱敏方式，要求系统有安全措施确保用户不能够绕过数据脱敏层次直接接触敏感数据。

6.4 选择脱敏方法

依据数据脱敏场景选择数据脱敏方法，数据脱敏方法参见附录A。

6.5 定义脱敏规则

依据已选择的数据脱敏方法，定义脱敏规则，并对常用数据脱敏规则进行固化，避免重复定义。

6.6 执行脱敏操作

脱敏操作需遵循个人隐私保护、数据安全保护等相关法规、行业监管规范或标准，个人敏感信息安全应遵循GB/T 35273中相关规定。根据已定义的数据脱敏规则，数据脱敏操作包括但不限于：

a) 对脱敏过程运行监控和分析；
b) 定期对脱敏工作开展安全审计；
c) 对脱敏任务自动化运行。

6.7 评估脱敏效果

在执行脱敏工作基础上，利用测试工具评估脱敏后数据对应用系统功能、性能等方面的影响，并根据验证情况不断优化脱敏规划。

附 录 A
（资料性附录）
数据脱敏方法

数据脱敏方法见表A.1。

表A.1 数据脱敏方法

序号	脱敏方法	方法描述	示例
1	掩码	用通用字符替换原始数据中的部分信息，掩码后的数据长度与原始数据一样。	将手机号码 13500010001 经过掩码得到 135****0001。
2	规整	将数据按照大小规整到预定义的多个档位。	将客户资产按照规模分为高、中、低三个级别，将客户资产数据用这三个级别代替。
3	替换	以虚构的数据代替真实的数据。	将姓名“张三”替换为“王二”。
4	乱序	对敏感数据进行重新随机分布，混淆原有值和其他字段的联系。	将金额 13526 乱序为 65123。
5	均化	针对数值性的敏感数据，在保证脱敏后数据集总值或平均值与原数据集相同的情况下，改变数值的原始值。	将 65、75、90、50 均化为 79、61、85、55。
6	散列	对原始数据取散列值，使用散列值来代替原始数据。	将 1234567 取散列值为 0100110。
7	数据截断	直接舍弃业务不需要的信息，仅保留部分关键信息。	将手机号码 13500010001 截断为 135。
8	日期偏移取整	按照一定粒度对时间进行向上或向下偏移取整，可在保证时间数据一定分布特征的情况下隐藏原始时间。	将时间 20180101 01:01:09 按照 5 秒钟粒度向下取整得到 20180101 01:01:05。
9	限制返回行数	仅仅返回可用数据集合中一定行数的数据。	商品配方数据，只有在拿到所有配方数据后才具有意义，可在脱敏时仅返回一行数据。
10	限制返回列数	仅仅返回可用数据集合中一定列数的数据。	查询人员基本信息时，对于某些敏感列，不包含在返回数据集中。
11	……		

参 考 文 献

[1] 全国信息安全标准化技术委员会等，大数据安全标准化白皮书（2018版）
[2] 贵阳大数据交易所，大数据交易区块链技术应用标准