匿名化“祛魅”——匿名化法律实务解析

作者：潘永建 | 朱晓阳 | 邓梓珊 | 左嘉玮

2021 年 8 月 20 日，《个人信息保护法》(“《个保法》”)正式颁布，并将于 2021 年 11 月 1 日正式施行。作为我国首部针对个人信息保护的单行立法，《个保法》对涉及个人信息处理的相关企业提出了更高的合规要求。作为例外，《个保法》第四条规定，个人信息“不包括匿名化处理后的信息”。换言之，若企业能对收集到的个人信息进行匿名化处理，其在后续的数据流通过程中无须承担额外的合规义务。因为《个保法》等相关法规未直接规定“匿名化”标准，企业对如何实现合格的“匿名化”普遍感到困惑。本文中，通力大合规团队在梳理匿名化、去标识化、假名化等概念的基础上，介绍常见去标识化技术，并解读已于 10 月 1 日生效的上海市地方标准《数据去标识化共享指南》(DB31/T 1311-2021)，对匿名化规则进行实务探讨。

一. 比较法视野下的匿名化

在计算机科学领域，数据匿名化通常指从数据集中删除个人身份信息(标识符)的过程。世界各国均认为此种处理可以有效减少隐私泄露的风险，从而在数据保护法律中对匿名化数据给予一定豁免，以平衡隐私保护和信息的自由流动。

在术语的使用上，各国存在交叉又有所区别。总体而言，我国的“匿名化”概念与欧盟的“匿名化”、美国的“去标识化”相似：是指个人信息经过处理无法识别特定自然人**且不能复原**的过程。经有效匿名化处理后的信息不再被认定为个人信息，不受个人信息保护相关法律规制。因为从理论上讲，个人信息主体已经无法被重新识别，没有必要强加额外的法定数据隐私保护。

如您需要了解我们的出版物，
请联系：

Publication@llinkslaw.com

我国的“去标识化”概念则类似于欧盟与美国的“假名化”概念：经过去标识化处理后的个人信息虽然无法直接重新识别到个人，**但可以借助额外的信息进行识别**。[1]因此，去标识化被认为是一种减少个人信息安全风险的手段[2]，但去标识化后的信息仍属于个人信息。至于我国的“假名化”概念，则是用于实现“去标识化”的一种常见技术。

相关概念归纳如下表：

概念\法域	中国	欧盟	美国
匿名化 anonymization	是指个人信息经过处理无法识别特定自然人且不能复原的过程 《个保法》第七十三条(四)	综合考虑技术、成本、时间等因素，如果数据控制者或其他人采用了所有合理可能(reasonably likely)的方法，仍无法直接或间接识别数据主体，则构成匿名化 GDPR Recital 26	
去标识化 De-identification	是指个人信息经过处理，使其在不借助额外信息的情况下无法识别特定自然人的过程 《个保法》第七十三条(三)		信息无法合理地、直接或间接地识别、关联到特定的个人，并且信息处理者通过技术措施、业务流程、自我承诺等方式确保其不会进行重标识 CCPA 1798.40(h)
假名化 pseudonymization	一种使用假名替换直接标识(或其他准标识符)的去标识化技术 《个人信息去标识化指南》附录 A.4	对个人信息的技术处理，使其在不结合额外信息的情况下，无法识别特定数据主体，且额外信息被分开存储并受技术、管理措施的保护 GDPR Article 4(5)	假名化是指不使用额外信息无法识别个人，且额外信息被分开存储并受技术、管理措施的保护 CCPA 1798.40(r)

在不同法域的数据匿名化制度中，美国更倾向于数据的合法流通，因此在技术认定标准上较为宽松，并未严格区分“去标识化”“假名化”。美国认为，信息处理者可以通过实施适当的组织措施以降低个人信息安全风险，从而满足监管要求。例如，根据美国联邦贸易委员会(FTC)发布的报告[3]，如果一家公司：

[1] 例如，通过加密技术实现的去标识化，可以通过密钥进行还原；通过 Hash 算法实现的去标识化，可以通过映射表单的方式还原。
[2] 参见《个人信息保护法》第五十一条。
[3] FTC, Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers, iv, 20-21 (2012)

(1)采取合理措施确保数据被去标识化; (2)公开承诺不会试图重新识别数据; (3)通过合同禁止数据接收者重新识别数据，那么数据就不能“合理链接”到个人(因此被排除在数据保护框架之外)。这一原则被之后的立法广泛采纳，包括《加州消费者隐私法(CCPA)》，《加州隐私权法(CRPA)》。

相较于美国法律，欧盟则更加注重于技术本身的有效性，只有(无论对于数据控制者或任何其他人)均无法直接或间接识别到数据主体的数据才构成匿名化数据，从而不适用 GDPR。但是，在考虑信息是否可以被重新识别时，欧盟也采用了“合理可能”的标准，其中检验“合理性”的因素包括所需的时间、成本以及现有技术等。假名化数据则仍然适用 GDPR，例如，欧盟第 29 条数据保护工作组(WP29)指出，假名化不是“匿名化的一种方法”，而仅是“降低了数据集与数据主体身份的联系”，因此可以作为一种有用的安全措施。[4]

我国《个保法》尚未具体规定匿名化的判断标准，根据文义解释，我国法律中的匿名化概念与欧盟 GDPR 较为接近。

二. 去标识化技术解析

(一) 基础概念

标识符	数据集中的一个或多个属性，可以实现对个人信息主体的唯一识别。
去标识化技术	降低数据集中信息和个人信息主体关联程度的技术。
重标识	把去标识化的数据集重新关联到原始个人信息主体或一组个人信息主体的过程。

(二) 常见技术手段

加密	通过密码学技术对标识符进行变换，包括确定性加密、保序加密、同态加密等。
抑制	对标识符进行删除或者隐藏，包括从数据集中删除整个标识符，或删除标识符的一部分。
泛化	降低数据颗粒度，使用概括、抽象的办法表示原有的数据项。
假名化	一种使用假名替换直接标识(或其他准标识符)的去标识化技术。 生成假名的方式包括加密或散列[5]等密码技术，实践中也可以多种技术同时运用以增加安全性。

[4] Article: 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques, Europa 3 (2014)
[5] 即 Hash 函数，一种加密技术，可将任意长度的二进制串映射为固定长度二进制串，常见的 Hash 算法有 MD5, SHA256 等。

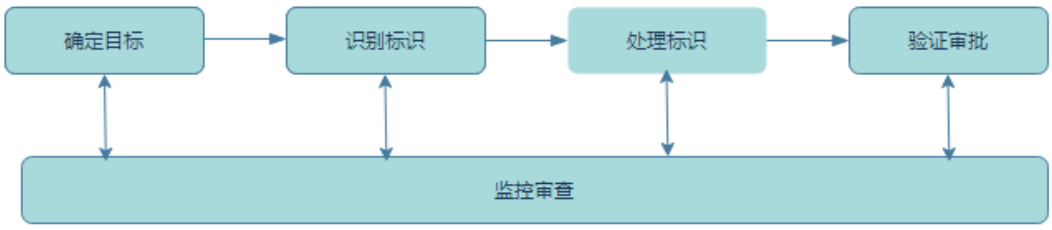
除了对标识符进行处理外，企业还可以结合数据应用场景选择合适的去标识化模型：

- K-匿名模型：要求发布的数据中，指定标识符属性值相同的每一等价类至少包含 K 个记录，使攻击者不能判别出个人信息所属的具体个体，从而保护了个人信息安全。
- 差分隐私模型：确保数据集中任何特定的个人信息主体的存在与否无法从去标识化数据集或系统响应中推导出。

(三) 去标识化流程

狭义上的去标识化通常指指对直接标识符和准标识符进行删除或变换，以避免攻击重新识别出原始个人信息主体；广义上说，去标识化还包括控制重标识风险、结合具体应用场景选择合适的去标识化模型和技术，并对去标识化效果进行评估的一系列过程。

因此，去标识化过程可分为确定目标、识别标识、处理标识以及验证审批等步骤，并在上述各步骤的实施过程中和完成后进行有效的监控和审查。如下图所示。



(四) 去标识化效果评估标准

根据《个人信息去标识化效果分级评估规范(征求意见稿)》("《分级评估规范》")，基于重标识风险从高到低，个人信息标识度分级可划分为 4 级：

重标识风险增大 ↓

4级	聚合数据
3级	重标识风险可接受数据
2级	消除直接标识符的数据
1级	能直接识别主体的数据

1 级数据：是指包含直接标识符的数据，在特定环境下能直接识别个人信息主体。

2 级数据：是指删除了直接标识符，但包含准标识符的数据。或者对直接标识符进行了处理(例如：泛化、抑制等)，使其不再能直接(单独)标识个人身份，并且重标识风险高于设定阈值(0.05)的数据。

3 级数据：是指消除了直接标识符，且重标识风险低于设定阈值(0.05)的数据。

4 级数据：是指对数据进行汇总分析得出的聚合数据，不再包含个例数据。例如总计数、最大值、最小值、平均值等。

遗憾的是，虽然《分级评估规范》给出了效果分级的精细标准，但仍未就“何种数据可以构成我国法律中的匿名化数据”这一核心问题给出明确答复，亦未给出推荐设定阈值 0.05 的计算依据。

三. 匿名化的风险与实践中的探索

(一) 匿名化的风险

从技术角度来说，如果给出足够的时间和资源，或许不存在绝对“不能复原”的数据，任何匿名数据都可能会被“去匿名化”——重新识别到个人信息主体。特别是考虑到数据挖掘技术的发展，以及日益增加的个人信息收集行为，重识别攻击将变得更加容易，去匿名化的风险也会增加。

例如，2006 年，Netflix 向公众发布了一个匿名的数据集，其中包括 50 万成员的电影评论。Netflix 对这些记录进行了“匿名化”处理，删除了用户名等识别信息，仅保留每个用户对电影的评分和评分的时间。发布后不久，两名研究人员发现，通过将 Netflix 评论与 IMDB 的公开数据进行匹配，可以重新识别 Netflix 数据集中大部分用户的身份。[6]

种种事例说明，尽管尽了最大努力去除可识别的信息，但理论上经过处理后的信息仍然存在被重新识别的风险，并且这种风险会随着技术的发展而持续增加，完美的“匿名化”技术可能仅仅是一种奢望。

(二) 实践中的探索

2021 年 7 月 27 日，上海市市场监督管理局发布了上海市地方标准《数据去标识化共享指南》(DB31/T 1311-2021)(“《指南》”)，并已于 10 月 1 日起实施。鉴于个人信息匿名化“不能复原”判断标准的模糊，且“绝对的不能复原”难以实现，《指南》对于数据去标识化共享采取了“相对的不可复原”标准，以平衡信息主体权利权益保护与数据流通。

《指南》承认纯粹的技术手段存在局限性、任何去标识化措施均存在被重新识别的可能，主张通过引入可信的第三方平台(例如数据交易中心)，在数据共享的过程中实施配套控制与保护措施，以防范个人信息安全风险，促进数据的合规流通。

[6] Arvind Narayanan, Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy

《指南》基础概念

数据标记	是数据标识(符)经去标识化处理后的结果，数据标记在数据共享域中具有替代标识(符)的索引作用。
数据共享域	以机构间数据共享为目的，由参与机构共同组织和运营的、有安全边界的、受控、互信且制衡的计算机域。
受控重标识	在数据共享域内，经评估合格后，在已知信息主体身份并获得同意的状态下，利用数据共享域的实时关联控制能力，将去标识化数据(集)重新关联到该信息主体或一组信息主体的过程。

《指南》数据流通规则

《指南》中的去标识化数据包括主体标识、数据项、数据值三要素，具体如下:

要素	描述	要求
主体标识	数据中信息主体的唯一性标识(符)	1、须对原始标识(符)进行去标识化处理，转换为无识别性的“标记”2、以“标记”为索引进行数据共享。
数据项	数据不同维度的定义(如数据库的各种字段定义)。	1、原始标识(符)不得成为数据项进入共享。2、除非评估认定具备个人明示同意的证据，数据项不得含有敏感数据的内容(敏感数据判定参见 GB/T35273—2020 附录 B)
数据值	对数据项的赋值，包括字符型、数字型、日期型、日期时间型、布尔型、二进制等类型	1、原始标识(符)不得成为数据值的部分或全部内容。2、数据值宜符合 K 匿名要求(除非具备个人明示同意的证据)。

根据《指南》规定，对于主体标识经处理后形成的“标记”，各主体原则上不得利用所接受的数据进行还原；对于其他去标识化后的数据，《指南》根据数据接收方的不同应用场景(群体或个体)，制定了相应的流通规则:

- 以群体为对象的应用(例如统计分析)，若不以识别特定人为目的，应当允许自由流通。**接收方可在维持数据去标识化的状态下，开展相关应用，无需就共享再次征得个人信息主体的同意**;
- 以个体为对象的应用(例如身份验真、贷款审批)，仍须遵循现行个人信息保护的法律规则，**在获得个人信息主体的同意之后，方可对去标识化后的数据进行“受控重标识”**。

并且，由平台方负责留存标记生成、受控重标识的记录，确保数据流通全过程可控和责任可追溯。

四. 结语

大量研究表明，目前几乎所有的匿名化处理技术均不可能达到 100%消除重识别风险的效果。因此，个人信息保护的目标应当是尽可能降低个人信息处理过程中的隐私风险，而不是追求技术上完美的匿名化。通过可靠的去标识化技术与配套的控制和管理措施相结合，未尝不是实现“匿名化”的一种途径。《指南》已经在地方层面对去标识化数据的共享进行了探索，期待可以推动相关国家标准的早日出台。

如您希望就相关问题进一步交流，请联系：

潘永建
+86 21 3135 8701
david.pan@llinkslaw.com

如您希望就其他问题进一步交流或有其他业务咨询需求，请随时与我们联系：**master@llinkslaw.com**

上海	北京	深圳
上海市银城中路 68 号 时代金融中心 19 楼 T: +86 21 3135 8666 F: +86 21 3135 8600	北京市建国门北大街 8 号 华润大厦 4 楼 T: +86 10 8519 2266 F: +86 10 8519 2929	深圳市南山区科苑南路 2666 号 中国华润大厦 18 楼 T: +86 755 3391 7666 F: +86 755 3391 7668
香港	伦敦	
香港中环遮打道 18 号 历山大厦 32 楼 3201 室 T: +852 2592 1978 F: +852 2868 0883	1/F, 3 More London Riverside London SE1 2RE T: +44 (0)20 3283 4337 D: +44 (0)20 3283 4323	

www.llinkslaw.com

Wechat: LlinksLaw

本土化资源 国际化视野

免责声明:
本出版物仅供一般性参考，并无意提供任何法律或其他建议。我们明示不对任何依赖本出版物的任何内容而采取或不采取行动所导致的后果承担责任。我们保留所有对本出版物的权利。

© 通力律师事务所 2021