数据匿名化的法律规制

王 融
中国信息通信研究院互联网法律中心 北京 100096

摘 要 数据匿名化是兼顾隐私保护与数据利用的有效手段。随着大数据、数据开放运动的发展，数据匿名化的法律问题受到更多关注。文章从法律视角对匿名化数据的概念、认定的法律标准以及数据匿名化处理过程中应当遵循的法律规范进行了系统性论述，提出匿名化区别于假名数据，对不可识别性有着更高的要求。对数据进行匿名化处理的机构，以及接收匿名化数据的第三方(在一般情形下)，均不能实现数据的身份再识别。为此，采取匿名化处理的机构应当在事前、事中、事后的整个周期采取必要的技术手段、合同机制以及IT审计等方式保障数据真正实现匿名化。

关键词 数据匿名化；数据保护法；法律规范

数据匿名化在计算机科学领域是方兴未艾的热门话题。自1997年美国学者Samarati和Sweeney提出k-anonymity匿名模型后，目前已发展出许多成熟的技术解决方案[1]。相比于技术领域的长足进步，法律领域对于匿名化的关注才刚刚开始(就国内研究看，通过中国知网查询“数据匿名”、“匿名化”，大量文献集中在计算机科学、通信科学领域，法学领域文献寥寥)。但随着全球个人数据保护立法制度的不断完善，关于数据匿名化的法律规范正逐步浮现。特别是在大数据浪潮的推动下，数据挖掘、数据共享(交易)、数据开放对于匿名化的需求越来越高，有关法律规范问题亟待澄清。本文从法律视角出发，对数据匿名化的法律规范问题予以系统分析。

1 数据匿名化的概念

如果从法律视角理解匿名化，则需要从个人数据保护法中的个人数据概念入手。一个基本的法律前提是：个人数据保护法规范个人数据的收集和处理活动，个人数据经匿名化之后不再适用个人数据保护法。

1.1 个人数据的概念

个人数据是个人数据保护法中的核心概念，其界定的关键是可识别特定自然人的数据，包括直接识别与间接识别。直接识别是指单凭该数据本身就能够指向特定个人，例如身份证号码、电话号码、家庭住址等；间接识别是指将该数据结合其他数据后也能够关联到特定个人。例如有关个人的一组常去地点数据，可以通过参考其他数据识别到特定个人。

相比于直接识别，间接识别标准为个人数据的界定预留了一定空间。特别是根据技术、应用的发展，可以通过立法等方式不断扩展个人数据的类别。例如：刚刚通过的欧盟《数据保护一般条例》(General Data Protection Regulation，GDPR)第4条中将位置数据(Location Data)、在线活动识别符号(Online Identifier，如IP地址、MAC地址、COOKIE等)明确纳入个人数据范围[2]。美国隐私保护最主要的执法机构美国联邦贸易委员会(Federal Trade Commission，FTC)在2012年发布的隐私保护指南，《在一个充满快速变化的时代，保护消费者隐私——关于隐私权的建议》中提出，FTC倡

导的隐私保护框架应当适用于：能够合理地与特定的消费者、电脑以及其他设备相联系的消费者数据的收集和使用行为。个人数据突破了传统定义中的与具体的自然人相关联，而是扩展到了用户所使用设备标识等[3]。这一扩展趋势在《美国儿童在线隐私保护法》执行规则中已经有所体现。规则中明确个人信息包括：能够被用来跟踪和识别个人在线活动的符号，例如IP地址、MAC地址等[4]。

因此而看，为应对大数据等信息技术的发展，各国通过立法适当扩展个人数据的范围，让在线跟踪、精准行为营销等新型数据处理活动也能够落入法律适用范围，以此对网络环境下的个人隐私保护提供新的补充。而在相反的方向上，匿名化则成为兼顾隐私保护和数据利用的另一条有效路径。

1.2 匿名化的提出

尽管人们在上世纪90年代就提出了匿名化思想，但大数据时代的到来真正让匿名化迅速成为热点技术。大数据中的大部分数据来源于人和传感器，包括用户上网浏览记录、社交网络上用户的信息、传感器数据和监视数据等。从浩瀚的数据宝藏中获得有价值的信息是各大企业收集数据的主要目的。数据成为企业最有价值的财产和新型商业模式的基石[5]。

与此同时，政府数据在促进创新方面所拥有的巨大经济、社会价值潜能也逐渐被各国所认知，自2009年起，在美国的引领下掀起了全球开放政府数据运动浪潮，推动以机器可读、可重复利用方式全面向社会公众开放。

然而无论是企业还是政府所掌握的数据中，都大量涉及个人数据，因此无可避免地遭遇隐私保护问题。个人数据保护和数据蕴藏的巨大价值之间的矛盾日益突出。匿名化作为二者的折衷，提供了一条技术解决路径。通过匿名化方法消除用户的身份信息、敏感信息以达到隐私保护的目的，同时还能够最大化地发挥数据价值。

匿名化技术发展的初衷主要是为了在数据利用的过程中降低个人隐私风险，因此它包含了不同的匿名化方法，例如泛化、压缩、分解、置换以及干扰等等，这些方法相应地负载不同程度的风险，有的相对容易被攻击复原，有的难度较高，甚至几乎不可能被复原。而从近年来各国个人数据保护法对于匿名化的理解看，法律语义下的匿名化相较于技术语义下的匿名化较为狭窄。根据个人数据保护法对个人数据界定的反向推导，法律语义下的匿名化数据至少要满足两个标准：1)仅从该数据本身无法指向特定的个人；2)即使结合其他数据也无法指向特定个人。

正如欧盟数据保护条例(GDPR)在引言部分对于匿名化的界定：“匿名化是指将个人数据移除可识别个人信息的部分，并且通过这一方法，数据主体不会再被识别。匿名化数据不属于个人数据，因此无须适用条例的相关要求，机构可以自由地处理匿名化数据。”也正是因为匿名化数据豁免了严格的个人数据保护规范，法律语义下的匿名化有着更高的标准——不可能被识别，或者被识别的可能性极低(为行文方便，下文中对于匿名化不可能被识别的定义包含了识别可能性极低的情况)。而在个人数据保护法下，与匿名数据相对应的另一个概念——假名数据，可被识别身份的风险相对较高。

1.3 与假名数据的区别

假名化(Pseudonymisation)是指对个人数据处理后，在没有特定信息参考(该特定信息被安全地单独保存)的情况下，不能指向特定个人。假名化数据与匿名化数据最大的不同是，前者仍然属于个人数据，仍要适用个人数据保护法。

换言之，假名数据仍建立在个体基础之上，保留了个体颗粒度，只是用了其他符号来标注数据，而非用真实的身份。假名数据结合了特定信息会恢复身份属性。例如：“张晓明，65岁，糖尿病患者”是真实的个人数据，在此基础上生成假名数据：“00108，65岁，糖尿病患者”。

因此法律要求，若要维持数据的假名状态，则机构必须单独安全地存储能够使得假名数据恢复身份数据的特定信息。例如，针对上面的例子，机构必须安全的存储“00108与张晓明”的对应关系信息。而匿名化数据在结合其他数据的任何情况下也不可能被识别。

相比于匿名化数据，假名数据的隐私风险会更高一些。但相比于真实的个人数据，假名数据有利于降低隐私风险，是帮助机构更好地履行数据保护义务的有效手段，因此在数据保护严格的欧盟，也鼓励机构对个人数据的假名处理。并且假名数据本身在科研领域有着重要价值，一方面假名数据最大程度地保留了信息价值，另一方面，科研领域数据公开的范围较小，仅限于科研人员使用，且有着较为完善的安全保障措施，隐私风险相对较低。

1.4 小结

个人数据不同的处理方法伴随不同的隐私风险，如果将数据看做是一段光谱，在光谱左端是真实的个人数据，其隐私风险最高，居中是假名数据，它保留了个体颗粒度，但隐私风险降低，右端是匿名数据(特别是匿名化后的聚合数据)，隐私风险努力接近于零。

2 匿名化的法律标准

如果缺乏其他的数据源，很多数据将保持匿名的状态。然而在大数据推动之下，有越来越多的数据集产生并公布，机构甚至是普通的个人都可以获取大量的数据资源。同时，软件算法和分析学的发展使得数据更易被关联和聚合，大大增强了人们将非个人数据转化为个人数据的能力。个人数据匿名化后遭遇有目的攻击的情况也更为普遍。

2006年10月， DVD租赁商奈飞公司开展“Netflix Prize”算法竞赛。该公司公布了大约来自50万用户的一亿条租赁记录，并且公开悬赏一百万美金，奖励工程师通过软件设计来提高其电影推荐系统的精准度。虽然奈飞公司对数据进行了精心的匿名化处理，然而其中部分数据仍然被认出，包括一名化名“无名氏”的同性恋，她因此起诉了奈飞公司[6]。在国内，某知名移动应用由于不注意保护用户位置大数据，攻击者可根据三角测量方法推断出用户的家庭住址等敏感位置[7]。

尽管数据匿名化的难度越来越高，但需重申的是：匿名化仍然是重要的数据安全保障措施，在大数据环境下，更应得到广泛的应用。匿名化是可能的，也是可行的。数据匿名化使得丰富的数据资源得以利用，同时也能最大程度保护个人隐私和数据。并且各国个人数据保护法对匿名化数据予以了法律适用上的豁免。那么，在法律上认可匿名数据需要考虑的因素是什么？如何确定匿名化的法律标准?

2.1 部分国家立法或监管机构对于匿名化的标准

1)美国。对于匿名化数据，法律中还没有明确细致的标准。但美国《健康保险可转移及责任法案》(HIPAA)对另一个相似的概念——去身份化(de-identification)作出了界定：“通过处理使得数据不能识别特定个人，或者没有合理的基础能够认为该数据可以被用来识别特定个人。”

2)日本。日本2015年通过《个人信息保护法》修正案，对于大数据交易做出修正规定。新法案允许企业向第三方出售充分匿名化的数据，但同时提出了相关义务要求：匿名后的数据不能够与其他信息进行比对、参照，以实现身份识别的功能，且不能复原。

3)新加坡。新加坡个人数据保护委员会2013年颁布的《个人数据保护法指定主题咨询指南》对个人数据的界定以及匿名化也作出了进一步规定。匿名化是指将个人数据转化成一种数据，这种数据无论是其本身，还是通过机构已经获得的或者可能获得的其他数据一起分析后都不能识别到个人。数据匿名化之后就不适用于个人数据保护法中的相关规定[8]。

4)英国。信息专员办公室ICO指出：匿名化并非是完全无风险的，而只是将风险降低到最小化。如果数据可被识别的风险是合理存在的，应当被视为个人数据[9]。

2.2 对匿名化标准的具体分析

从各国数据保护立法和监管实践看，匿名化数据的法律标准有以下几个共同要素。

1)对于数据控制者(data controller，是个人数据保护法中的主体概念，指单独或联合其他方决定了数据处理目的和方法的主体，如公司 、公共机构等)来说，匿名化的数据不能够再被识别特定个人。如果对数据进行匿名化处理的机构仍然掌握着恢复该数据身份属性的关键信息、算法，则对于该机构来说，这些数据仍然属于个人数据，仍需要适用个人数据保护法。因为对于该机构而言，其随时可以恢复该数据的身份属性，不属于不能识别的情形。

2)如果数据匿名化之后被公布，则除了数据控制者之外，还要考虑能够获得该数据的第三方，其所能够获得的数据资源、数据能力，以及在此基础上对匿名化数据重新恢复身份的可能性。欧盟个人数据保护的指导机构——第29工作组，关于个人数据概念的法律意见表明(EU Article 29 working party Opinion (4/2007) on the concept of personal data)，匿名化数据不能识别出特定个人是指：不论是对于掌握该数据的数据控制者，还是其他获得该数据的任何主体，采取一般可能的措施、手段也无法将该数据关联到特定个人。

3)依照第2)点标准带来的一个现实问题是：对于向社会公开的匿名化数据，接收匿名化数据的第三方的数据能力是千差万别的，其所能获取的数据资源也各有不同。因此对于泛泛主体而言，数据是否匿名的评判结果是因人而异的。

针对这一问题，在技术领域提出的“有动机的攻击者测试”(a motivated intruder test)也逐步为监管实践所认可。例如英国数据保护机构ICO推荐在判断匿名化问题时应用该测试[10]。该测试假设包括以下几个方面。第一，有目的明确的入侵者，他没有预先掌握相关的其他资源，只是希望从公开的匿名数据集中重新识别数据。匿名数据被重新识别，并不仅仅指该数据所指向的个人的名字被知道，只要能够建立该数据与特定个人的可靠联系就可以被认为是“被识别”。第二，该入侵者具备一般人的合理能力，能够获得相关资源(例如包括互联网、图书馆、公开文件等)，但通常不能将该入侵者假设成为具备特殊技能知识的专家(如黑客)，或者是极端情况下的犯罪分子(如入室非法获取数据)。因此，该测试的可应用性在于它比普通大众的标准要高，同时又比专业人士的标准要低，在一个合理可能的范围内。

应用该测试，考虑数据是否匿名的主要因素包括以下3点。

1)动机。即尝试重新识别的可能性。在实践中，一些特定类型的数据对于有动机的入侵者来说更有吸引力，对个人带来负面结果的可能性也就越大。例如能够通过去匿名化获得经济利益或其他不法利益，或者去匿名化后能够引起刺探他人隐私、引起他人尴尬的数据。

2)可获取的其他数据资源。包括通过政府数据开放能够为社会大众所获取的数据，也包括通过共享、交易方式获得的其他数据。

3)技术能力。对技术能力的考虑应当结合当下技术的最新发展。特别是在数据分析、挖掘技术快速发展的今天，当前的匿名化并不代表永久的匿名化。

总之，尽管立法和监管实践竭尽努力地为数据匿名化提供尽量客观的判断标准，但实际上，对数据重新识别身份的风险给出绝对结论依然是不严谨的。具体案例具体分析仍然是判断匿名化问题的第一法则。当然，以上所列出的具体标准分析可以作为实践的重要参考。

3 数据匿名化的法律规范

匿名化数据不受个人数据保护法保护是本文一再重申的观点。这一观点主要表达的是：数据经充分匿名化后，数据控制者对于该数据的使用处理不再受个人数据保护法的规范，例如包括知情同意原则、目的限制原则、最小化原则等都不再发挥约束作用。并且由于这些数据切断了与特定个人的联系，数据控制者也无需为个

人数据权利(如知情权、访问权、拒绝权、删除权)的实现提供支撑。

然而，享受个人数据保护法的豁免待遇，则需要为此承担其他的法律义务，这些义务一方面来源于个人数据保护法对于匿名化数据认定的高标准(即通过施加此类义务，实现真正的数据匿名化)，另一方面来自于信息安全其他方面的法定要求。我们可以将这些规范分为事前、事中、事后三个阶段。

3.1 事前阶段

1)关于同意。开展数据匿名化，数据控制者面临的第一问题——这是否需要经得用户的同意。一般情况下，匿名化处理个人数据并不需要征得用户同意。单纯的匿名化，是有助于数据安全的有效手段。在这个阶段，征求用户同意将会十分繁琐，甚至不可行。但是这并不意味着对匿名化没有透明度方面的要求。相反，如果机构能够通过隐私政策在事前告知用户数据可能的匿名化利用，则可以作为最佳市场实践推荐。例如：无论是手机制造商还是应用开发商在使用用户位置信息时，应当向用户明确告知其个人位置信息是如何被使用的，是以个人数据模式使用，还是以匿名化的方式被使用。

值得注意的是，如果机构收集非个人数据，但之后利用某种手段或者技术方式，再次恢复数据的可识别性(re-identification)，而这种收集与识别活动并没有得到用户知情或者同意，则涉嫌违反个人数据保护法。

2)隐私风险评估。在数据匿名化的初始阶段，开展隐私风险评估非常重要。特别是如果数据匿名化的目的是将数据开放给公众，或者与其他主体进行共享、交易，则可能会产生重新被识别风险。若一旦被识别，即使后续采取补救措施，其影响也是不可逆的。因此要充分结合匿名化标准的三个要素进行充分的评估，包括有动机的攻击者，其能掌握获得的其他数据资源，以及可以被利用的去匿名化技术等。根据隐私风险评估的结果，机构可以选择不同的加密方式和利用方式，以便有针对性地消减隐私暴露风险。

3.2 事中阶段

1)根据隐私风险评估，对匿名化策略做出调整。例如：选择更为复杂的匿名化方法，缩减信息披露的规模，限制披露的对象，附加合同约束条件等等。一般情况下，匿名数据越具体、越接近个体颗粒度状态，则在披露时越需要限定在一定范围，限定一定条件。越聚合化的信息，则更倾向于完全公开的方式，比如犯罪热点地图。

数据控制者必须意识到，限定在一定范围、程度内的匿名化披露，需要有强有力的安全保障措施，包括但不限于以下几点。

①目的限制， 数据仅仅能够服务于双方约定的目的；

②对匿名数据接受者的员工进行安全培训，特别是数据访问方面要符合安全原则和最小化原则；

③对将其他数据带进使用场景的能力进行控制，从而对应用关联数据方法实现身份识别的风险进行管理；

④限定数据仅能在特定的项目中使用；

⑤限制对匿名数据的再披露；

⑥禁止尝试对数据进行重新再识别，并且对于随机情况下再识别的数据做出破坏处理；

⑦加密以及秘钥管理以限制对数据的访问；

⑧在项目完成之后对数据的返还、销毁做出协议安排；

⑨责任追究机制，即数据接收方一旦违反合同协定需要承担违约责任。

2)持续的隐私风险评估。隐私风险评估不仅在事前阶段需要，它还将贯穿匿名化数据利用整个过程。重新识别的风险会随着时间的推进而变化。一方面数据分析技术在飞速发展，过去公布的匿名化数据未必在新的技术、新的模型面前仍然保持匿名；另一方面，政府数据开放，以及私营主体之间的数据交易、共享使得新的数据集在源源不断的被释放，这些都需要机构应当定期评估数据匿名化策略，特别是对于定期例行公布匿名数据

的机构来说(因为受到政府信息公开或者数据开放的要求，这在政府或公共机构更为常见)，应定期开展隐私风险评估，明确当前及可预见未来的威胁，并及时对数据匿名策略作出调整。

3.3 事后阶段

数据匿名化事后阶段的核心任务是始终保持数据的匿名状态。数据利用限定在非身份化的模式之下，在后续的利用中也不再进行身份识别，企业可通过合同机制、IT审计等方式予以监督。在后续利用阶段，企业若将匿名数据恢复身份属性，则数据处理活动再次落入到个人数据保护法调整范围内。此外，一旦在攻击事件中被去匿名化，机构应当启动应急预案，包括通知受到影响的个人，并帮助其采取必要的补救措施。

4 结语

数据匿名化不能仅仅被看作是脱离于数据保护法之外，避免管制负担的一种手段。应用它的初衷是降低个人数据泄露的隐私风险。采取匿名化措施的企业能够向用户提供更多的安全保障，让用户知晓其被收集的信息在用于大数据分析时，并没有使用可识别身份的数据，因此增强用户对大数据应用的信任和安全感。为保证匿名化更多地发挥安全屏障作用，而不是作为数据滥用的挡箭牌，匿名化利用应当在合法合规的前提下开展。

2016年7月5日，《中华人民共和国网络安全法(草案二次审议稿)》正式向社会公布。与一审稿相比，草案增加了类似匿名化的规定。例如“第四十一条：网络运营者不得泄露、篡改、毁损其收集的公民个人信息；未经被收集者同意，不得向他人提供公民个人信息。但是，经过处理无法识别特定个人且不能复原的除外”。这里的特殊规定，可以理解为对于个人数据匿名化利用，特别是匿名化后对外提供(交易)的情形提供了合法性。在此基础上，建议我国应当加快建立数据匿名化利用的法律规范体系，包括：明确匿名化数据的法律概念和认定标准，强调数据不再具有身份可识别性；引入隐私风险评估机制，鼓励企业基于个案在内部实施数据匿名化的风险评估，并基于评估结果，适时调整匿名化策略；利用合同规范、技术保障等多重工具实现数据的真正匿名化；建立数据匿名化的事前、事中、事后规范体系。

参考文献

[1] 丁旋.社交网络分析中的隐私保护问题:去匿名化与无缝隐私[D].清华大学,2014

[2] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) [R].THE EUROPEAN, 4.5.2016 L 119/1 Official Journal of the European Union

[3] FTC REPORT: Protecting consumer privacy in an era of rapid change, recommendations for businesses and policymakers[R/OL].[2016-08-02].https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers

[4] Complying with COPPA: Frequently Asked Questions[EB/OL].[2016-08-02]. https://www.ftc.gov/tips-advice/business-center/guidance/complying-coppa-frequently-asked-questions

[5] 刘雅辉,张铁赢,靳小龙,等.大数据时代的个人隐私保护[J].计算机研究与发展,2015, 52(1):229-247

[6] 维克托·迈尔-舍恩伯格,肯尼斯·库克耶.大数据时代:生活、工作与思维的大变革[J].盛杨燕,周涛,译.杭州:浙江人民出版社, 2013

[7] 王璐,孟小峰.位置大数据隐私保护研究综述[J].软件学报,2014,25(4):693-712

[8] Singapore Personal Data Protection Commision: Advisory Guidelines on the Personal Data Protection Act for Selected Toptics Issued by the Personal Data Protection Commision[R]. Issued 24 september 2013, revised 16 May 2014, revised 11 September 2014

[9] UK. Information Commissioner's Office. Big data and data protection[R]. 2014

[10] UK. Information Commissioner's Office. Anonymisation: managing data protection risk code of practice[R].2012

作者简历

王 融

中国信息通信研究院互联网法律中心副主任，高级工程师。长期从事电信、互联网立法与监管政策研究。代表著作(合著)：《电信法》、《融合背景下的中欧电信管制比较研究》、《个人信息保护法研究》。近年来主要研究方向为个人信息保护法、网络信息安全法。发表论文、文章40余篇，重点支撑信息通信领域法律、行政法规及部门规章立法工作。

The legal Requirements for Data Anonymization

Wang Rong Internet Law Center of China Academy of Information and Communication Technology, Beijing 10096, China

Abstract Data anonymization is an effective means which could compromise the privacy protection wih the use of data well. The legal issues on the data anonymization are getting more and more attention with the development of the big data and data open initative. From the legal perspective, this paper demonstrates the legal concept of anonymous data, the judgement standard of it as well as the legal requirements on the process of data anonymization in details. It indicates the anonymous data is different from pseudonym data, the former of which has a higher requirement for the data being unidentifiable status. The organization that anonymize the data could not make the data re-identifiable. None of the third party which receive the anonymous data could do that under the general circumstances. In order to achieve this objective, the organization should take all necessary means including the technical measure, contract mechanism and IT audit to ensure that data being unidentifiable in the whole process.

Keywords Data Anonymization; Data Protection Law; Legal Requirements

（上接15页）

Telecom Operators' International Roaming Service Promotion Based on Big Data

Xia Chao China Unicom, Tianjin Branch, Tianjin 300052, China

Abstract With the popularity of 4G networks and smartphones, the demand of international data roaming increases. International roaming service promotion based on big data not only meets the users' demand, but also makes full use of modern information technology to improve the enterprises' marketing efficiency. This article introduces the development and opportunities of telecom operators' international roaming service, the trends of big data and cloud computing application, and emphasizes the necessity of international roaming service promotion based on big data, and describes the promotion mode. Telecom operators, as the medium of the users' everyday communication, can get various kinds of real-name user data based on real usage behavior. Through the use of big data technologies, telecom operators could integrate data more effectively, and select target users for marketing, explore the value of big data and create benefits for the enterprise.

Keywords Telecom Operators; Big Data; Cloud Computing; International Roaming; Service Promotion