实施数据去标识化的指南

FUTURE OF PRIVACY FORUM
由 FPF.ORG 制作

EY

科学家、监管部门和律师是如何理解去识别化的？去识别化数据与匿名数据或假名数据有什么区别？数据可识别性不是二元对立的。它反而属于多种层级。

这是有关如何区分不同类别数据的指南。

可识别性的层级

包含直接与间接标识符的信息。

假名数据

直接标识符消除化或转变化的信息，但间接标识符还留存。

去识别化数据

直接和已知间接标识符被消除化或操控化，故此打破与现实身份的联系。

匿名数据

直接和间接标识符都通过数学和技术机制移除或修改，以此防止被重新识别。

直接标识符

直接识别个人身份的数据。无需额外信息或与公共领域信息进行连接（例如姓名、社会保障号码）。

间接标识符

以间接的方式识别个人身份。有助于连接信息片段，直到可以挑出一个人为止（例如生日、性别）。

保障和控制措施

技术、组织和法律上的控制措施。防止员工、研究人员或其他第三方进行重标识。

	明确个人	可识别化	不方便识别化	密钥编码	假名	受保护假名	去识别化	受保护而被去识别化	匿名	聚合匿名
直接标识符	完整	部分隐藏	部分隐藏	消除化或转变化	消除化或转变化	消除化或转变化	消除化或转变化	消除化或转变化	消除化或转变化	消除化或转变化
间接标识符	完整	完整	完整	完整	完整	完整	消除化或转变化	消除化或转变化	消除化或转变化	消除化或转变化
保障和控制措施	不相关 - 由于数据的性质	有限或无控制机制	具有控制机制	具有控制机制	有限或无控制机制	具有控制机制	有限或无控制机制	具有控制机制	不相关 - 由于数据的性质	不相关 - 由于高度数据聚合
例子	姓名、地址、电话号码，身份证	设备标识符、车牌、病历编号、cookie、IP地址 (例如, MAC 地址 68:A8:6D:35:65:03)	相同于可识别化数据，除了受保障和控制的数据（例如，散列MAC地址与法律陈述）	保管只能存取临床或研究集（例如王先生，糖尿病，HgB 15.1 g/dl = Csrk123）	独特人工假名替换直接标识符 (例如王先生 = 5L7T LX619Z) (独特序列没有其它用处)	跟假名一样除非保障控制的机构也保护数据	数据被压制，广义化，扰动，交换等 (例如 GPA:3.2 = 3.0-3.5、性别: 女性 = 性别: 男性)	跟去识别化一样除非保障控制的机构也保护数据	例如将数据集用于校准噪音，故此隐藏某个人是否在场（差分隐私）	极高度聚合数据（例如统计数据、人口普查数据、或人口数据——显示 52.6% 的华盛顿州 (DC) 居民是女性之类的）