The Seven States of Data

Khaled El Emam[1,2], Eloise Gratton[3], Jules Polonetsky[4], Luk Arbuckle[5]

[1] University of Ottawa

[2] Privacy Analytics Inc. (part of IMS Health)

[3] Borden Ladner Gervais LLP

[4] Future of Privacy Forum

[5] Children's Hospital of Eastern Ontario Research Institute

There has been considerable discussion about the meaning of personal information and the meaning of identifiability. This is an important concept in privacy because it determines the applicability of legislative requirements : data protection laws around the world protect and govern personal information.

There is a general view that identifiability falls on a spectrum, from no risk of re-identification to fully identifiable, with many grades in between. This spectrum has been defined previously in terms of a probability. Recently, a number of legal scholars have proposed different approaches to determine at what point information should be considered as "personal information", in many cases, using a risk based approach. For instance, Schwartz and Solove define three specific states of data: identified, identifiable, and non-identifiable. There is also the on-going distinctions between de-identified and pseudonymous data, whereby the latter is considered personal information by regulatory authorities.

In this article we examine these states of data on the spectrum of identifiability in more detail, and characterize the manner in which they differ. These identifiability states of data are colored by our experiences with health data, although they may nevertheless be useful much more broadly to other domains.

The objective of our analysis, in addition to being descriptive, is to formulate criteria for the sharing of pseudonymous data. These criteria are consistent with existing notions of re-identification risk and therefore do not require changes to the current framework for defining identifiability.