
IV
JORNADA PROFESIONAL DE LA RED
DE BIBLIOTECAS
DEL INSTITUTO CERVANTES:
**Big Data y bibliotecas: convertir**
**Datos en conocimiento**

**MADRID 11 DE DICIEMBRE DE 2014**

# Desafíos que para la privacidad y la protección de datos implica el Big Data

**Judith González Pedraz**
Asesora de la Unidad de Apoyo al Director
Agencia Española de Protección de Datos

Una de las cuestiones que está siendo objeto de estudio pormenorizado por la Agencia Española de Protección de Datos en estos momentos es el fenómeno del Big Data y sus implicaciones en materia de protección de datos.

*Big Data* o *Datos Masivos* es un término que hace referencia al enorme incremento en el acceso y uso automatizado de información. Se refiere a las gigantescas cantidades de datos digitalizados que son controlados por las empresas, autoridades públicas y otras grandes organizaciones que poseen la tecnología para realizar un análisis extenso de los mismos basado en el uso de algoritmos.

Así lo define el Documento de Trabajo sobre *Big Data and Privacy* elaborado por el Grupo Internacional de Trabajo sobre Protección de Datos en el sector de las Telecomunicaciones (mayo de 2014), grupo del que formamos parte como autoridad de protección de datos.
El término Big Data o *datos masivos* lleva relativamente poco en circulación -hace no más de un par de años que comienza a hablarse del mismo- y son la digitalización y las nuevas técnicas de procesamiento de datos los que lo ha hecho posible. Supone una revolución en el modo de recopilar y analizar los datos.

Tres son las características que definen al Big Data: *volume, variety* y *velocity.* Es lo que se conoce como las *tres uves* de Big Data. Se trata de una gran cantidad de datos de diferentes tipos o categorías que se generan y tienen que ser tratados a gran velocidad. Entre esos datos se encuentran innumerables datos personales.

Ha aumentado enormemente la escala de los datos que se manejan con lo se pueden hacer cosas que antes no era posible. Nunca hubiéramos imaginado que sería posible medir, almacenar, analizar y compartir estas vastas cantidades de datos. "*En el 2013 se estima que la cantidad total de información almacenada en el mundo es de 1.200 exabytes, de los que menos del 2% no es digital (…) Si estuvieran impresos en libros cubrirían la superficie entera de EEUU, formando unas 52 capas"* (Viktor Mayer-Schönberger y Kenneth Cukier en "La revolución de los datos masivos").

Los datos masivos serán una fuente de innovación y de nuevo valor económico. Pueden tener aplicación en una gran variedad de ámbitos, no sólo para definir hábitos de consumo o crear perfiles de consumidores, sino en el ámbito de la seguridad nacional, la investigación científica, los estudios médicos, prevención de catástrofes naturales o de propagación de enfermedades o epidemias, persecución del fraude fiscal, etc.

Pero estos avances entrañan también importantes riesgos para la privacidad de los ciudadanos que deben ser valorados con el fin de garantizar su derecho a la protección de datos personales.

Esos datos masivos pueden ser empleados por los bancos, compañías de seguros u otras grandes empresas y también por las autoridades públicas para tomar decisiones que afectan de manera determinante a la vida de esos individuos de los que se han creado perfiles.

Piénsese, por ejemplo, que nos denegasen el acceso a un seguro médico porque según determinados parámetros somos propensos a padecer cierta enfermedad cuyo tratamiento resultará muy costoso o que nos deniegan la entrada en un país porque estamos identificados como "potenciales delincuentes" según nuestro perfil psicológico y sin que hayamos cometido ningún delito.

Como vemos, los datos masivos se emplean para hacer predicciones (basándose en correlaciones, es lo que se denomina análisis predictivo) y tratan del *qué está ocurriendo o va a ocurrir*, no del *por qué se ha producido o va a producir*.

La era de los datos masivos puede llegar a alterar el funcionamiento de los mercados y la sociedad conduciéndonos a la llamada *dictadura de los datos* (Viktor Mayer-Schönberger) y supone un reto para la Agencia anticiparse a las demandas de los ciudadanos para salvaguardar su privacidad en este nuevo contexto.

El uso de Big Data supone un reto en materia de aplicación y cumplimiento de la normativa de protección de datos (Directiva 95/46/CE, de 24 de octubre de 1995, sobre protección de datos y la legislación nacional: Ley 15/1999 y su reglamento de desarrollo del año 2007).

Aunque algunos afirman que la legislación actual en esta materia se ha visto superada por el fenómeno de Big Data y debe modificarse para adaptarse al mismo, considero que en este contexto es más importante que nunca mantener y salvaguardar firmemente los principios fundamentales en materia de protección de datos que constituyen una garantía incuestionable para el derecho fundamental de los ciudadanos.

Los principios clave cuyo respeto y observancia es especialmente relevante en el ámbito de Big Data serían:

- Principio de legitimidad y consentimiento: para que el tratamiento del dato personal sea legítimo el afectado ha de prestar su consentimiento inequívoco.
- Principio de limitación de la finalidad: los datos han de ser utilizado sólo para la finalidad para la que fueron recabados.
- Principio de calidad: los datos han de ser adecuados, pertinentes, no excesivos, exactos y actualizados.
- Principio de minimización de los datos: que exige utilizar sólo los datos estrictamente necesarios para cumplir el fin para el que se recaban y no más.
- Principio de información o transparencia: derecho del ciudadano de conocer y acceder a toda la información que se posea sobre el mismo.

La legislación de protección de datos no se opone al desarrollo y aplicación del Big Data pero este fenómeno debe implantarse partiendo siempre del respeto a estos principios. Esto no obsta para que la normativa tradicional en materia de protección de datos pueda y deba verse completada con otras aproximaciones con el fin de salvaguardar la privacidad de manera efectiva. De ahí la importancia de realizar evaluaciones de impacto en la protección de datos y de potenciar la transparencia por parte de las empresas que tratan datos masivos, así como de otorgar a los ciudadanos un adecuado control sobre sus datos personales y de reforzar la necesidad de que otorguen su consentimiento.

Los **desafíos** que el *uso de Big Data* abre en el ámbito de la protección de datos personales -y que estamos estudiando- no son pocos, **problemas de** obtención del consentimiento, de ejercicio de los derechos de información y acceso, rectificación, cancelación u oposición, la cuestión de la correcta anonimización de los datos antes de analizarlos o la del desvío de la finalidad para la que fueron recogidos y otras posibles vulneraciones del principio de calidad de los datos (datos inexactos o excesivos).

1º) Los datos personales van a ser reutilizados para una **finalidad** diferente de aquella para la que fueron recogidos. Esto exige recabar previamente el **consentimiento** de los ciudadanos antes de realizar ese análisis predictivo. Las empresas deberán informar de que los datos personales van a ser utilizados para ser analizados de determinada forma, por ejemplo, con fines comerciales, para establecer perfiles de consumidores. Lo cual resulta muy difícil en la práctica cuando son terceros distintos de los que recogieron los datos quienes van a analizarlos.

Nos preguntamos cómo se van a dar cumplimiento en este contexto a los deberes de **información** en la recogida de datos y de obtener el consentimiento del afectado para esos nuevos tratamientos, que, además, en la mayoría de los casos se desconoce a priori cuáles serán.

Cuando no sea posible recabar el consentimiento será preciso asegurar la anonimización de los datos con todos los problemas que conlleva lograr una anonimización efectiva, que no sea reversible.

2º) El fenómeno del Big Data supone recoger y almacenar más datos de los estrictamente necesarios para cumplir la finalidad para la que se recogieron. Los datos adquieren un valor en sí mismos por los potenciales futuros usos que puedan tener. Por tanto, nos encontramos con **datos** que resultan **excesivos** en ese aspecto y plantean también un problema de cumplimiento del principio de calidad de los datos.

3º) Implica también que los **datos no** van a ser **cancelados** cuando dejen de ser pertinentes para la finalidad para la cual hubieran sido registrados. Esos datos masivos se conservan permitiendo la identificación del interesado y las autoridades de control tenemos que exigir su anonimización o su cancelación. Se corre el riesgo de que miles de datos personales anden en circulación y se compren y vendan para fines que nada tienen que ver con aquellos para los que fueron recabados y para los que los ciudadanos prestaron su consentimiento.

4º) Dificulta el ejercicio del **derecho de acceso a la información** que las compañías poseen sobre nosotros. La falta de transparencia y de información sobre cómo se almacenan y usan nuestros datos hace que seamos víctimas de decisiones que desconocemos y no podemos controlar. La mayoría de usuarios de Internet desconocen que los datos que se recopilan en su navegación son empleados para dirigirles publicidad comportamental, por ejemplo.
Esto adquiere mayor gravedad si mediante la combinación de datos procedentes de distintas fuentes puede obtenerse información sensible de una persona como su ideología, orientación sexual o estado de salud. No hay que olvidar que estos datos están especialmente protegidos por nuestra legislación.

Los ciudadanos deben ser informados sobre qué datos personales son recogidos, cómo van a ser tratados, para qué finalidades serán usados y si serán cedidos a terceros. Además deben poder conocer sus perfiles y acceder a la información a la que se aplican esos algoritmos para desarrollar los perfiles. Eso implica conocer también las fuentes o bases de datos de las que se han sacado sus datos.

En este ámbito los ciudadanos deberán tener derecho a que el responsable de los datos les entregue en un formato adecuado para su lectura y portabilidad todos los datos que posee, según el documento del Grupo de Berlín antes citado.

5º) Existe el riesgo de que las decisiones basadas en Big Data partan de **datos** que sean **inexactos**. La información o los datos que se cruzan para realizar el análisis proceden de diferentes fuentes, no todas son registros públicos y no son verificados, por lo que pueden no ser correctos. Además estamos hablando de datos que se han recogido para otro propósito y generado en un contexto distinto a aquél al que van a ser aplicados por lo que los resultados que se obtengan en muchas ocasione no van a

reflejar la situación actual. Los datos masivos no tienen en cuenta el contexto que es un elemento importante.

Por ello, es necesario que los ciudadanos tengan acceso a la información que se dispone sobre ellos para poder ejercer su **derecho de rectificación o cancelación** de aquellos datos personales que sean inexactos.

6º) Finalmente, destacar el riesgo de **re-identificación** que supone el uso de Big Data. Los individuos podrán ser identificados con relativa facilidad mediante el análisis de **información** procedente de diferentes bases de datos y que en apariencia está correctamente **anonimizada**.

Hay que tener en cuenta que mediante el empleo de *datos masivos* un ciudadano puede ser identificado no sólo mediante una identificación directa con los datos provenientes de una única base de datos, sino a través de la combinación de datos procedentes de múltiples bases de datos, ya sean públicas y/o privadas.

Así, una empresa puede creer que tiene correctamente disociados los datos que maneja pero desconoce si hay otros terceros que van a manejar sus datos (por ejemplo, para hacer perfiles de consumidores) que poseen la capacidad de acceso a otras bases que contienen mucha más información, la cual analizada de determinada manera puede permitir re-identificar a esos consumidores iniciales.

Por ello, sería necesario que las empresas y autoridades que manejen datos masivos realizasen una **evaluación de impacto en la protección de datos** para ver qué riesgo existe de re-identificación. Esto es, aplicar la *privacidad desde el diseño*. En esa evaluación se tiene que plantear qué otros datos sobre esos ciudadanos están o pueden estar disponibles, ya sea en fuentes de acceso público o bases de datos de otras empresas y qué posibilidades existen de que puedan ser puestas en relación con éxito logrando esa re-identificación. Esto dependerá de las técnicas para anonimizar los datos que se hayan aplicado.

En este sentido, el Grupo de Trabajo formado por las autoridades europeas de protección de datos, el llamado Grupo del artículo 29, ha emitido el Dictamen 5/2014 sobre las técnicas de anonimización de datos personales. No obstante, la constante evolución de las nuevas tecnologías hace cada vez más difícil que podamos pensar en una anonimización irreversible. Por ello, habrá que estar atentos en la salvaguarda de esos datos que no dejan de ser personales si puede volver a identificarse el ciudadano al que corresponden. Como vemos, los retos que nos plantea el Big Data van a exigir unas autoridades de control activas y vigilantes que den respuesta a la exigencia de protección de su privacidad que nos van a plantear los ciudadanos.