HACIA UN NUEVO DERECHO EUROPEO DE PROTECCIÓN DE DATOS

TOWARDS A NEW EUROPEAN DATA PROTECTION REGIME

Artemi Rallo Lombarte
Rosario García Mahamut

Editores

Valencia, 2015
Data origin and the proposed regulation

Martin Abrams

Executive Director
The Information Accountability Foundation

1. INTRODUCTION

Excellent public policy begins with a deep understanding of the problem to be solved. The European Union is in the midst of reforming data protection law. The institutions working on the proposed data protection regulation have a deep understanding of the privacy values that data protection must protect. However, it is less certain that they understand the data they are attempting to regulate.

In the simplest terms, data protection requires that data be collected, stored and used, and discarded when it no longer is used for the purposes for which it was collected, in a lawful and fair manner. Data protection law is very much based on this linear progression that runs from collection to destruction. Furthermore, data protection works best when the data collection is visible to the individual to which it pertains.

But what if there is a basic flaw in the lifecycle? What if data, rather than being collected, is instead created from other data? In addition, what if that data is then used to create even more data? There is clear evidence that the questions just posed are true. To establish that case, this paper first lays out a data taxonomy based on how the data originates. This origination taxonomy was created by the Information Accountability Foundation for a 2014

---

1 For more than 35 years, Abrams has been an information and consumer policy innovator. His most recent work has examined Big Data governance and privacy compliance driven by demonstrable data stewardship.
OECD experts meeting on privacy and big data. Sections of that paper are incorporated. The paper will then look at how well the proposed regulation handles the classification system described.

2. TAXONOMY

Data constitutes the lifeblood of an information age by forming the basic building blocks of all business, government and social processes. As data growth accelerates, much of it pertains to individuals either directly or indirectly. For example, data generated by the sensors in our tires links to the vehicle which, in turn, links to the car’s driver. In addition, more and more of that data is addressable by analytics processes. Those processes drive innovation and create economic and social value. They also create risks that individuals will be harmed in some tangible, inappropriate fashion or that individual dignity will be impacted in a fashion society considers unfair. To both facilitate innovation and protect individuals, data and its uses must be governed. Governance must be effective given the true nature of data in 2014 and beyond.

Along with the growth in data has come a fundamental change in the data itself. The computerized systems that inspired legacy privacy guidance was mostly contributed by individuals directly as those individuals participated in commerce and other facets of life. Today, more and more data originates from observations that are less obvious to the individual and are a product of processing itself. These new data will only increase as society builds out a more sensor-rich environment, and organizations make greater use of advanced analytic processes like Big Data.

2.1. Background

Collection has been the nexus for governing data since the publication of Privacy and Freedom by Alan Westin in 1967. Westin’s work, along with the work of other scholars established a road map
for protecting privacy when societies were in the early stages of automating information that pertains to people. The early scholarship established the contextual nature of privacy and suggested individual control the best means for governance. Early laws and guidelines put individual control in place through notifications of collection and purpose, in addition to individual consent for the listed purposes. Further, governance guidance was designed to be supportive of the control that comes from participation in data creation. The nexus for governance according to that model is the collection of data from the individual. The taxonomy in this paper will refer to that data type as provided, since the individual provides the data as part of interaction with the user (often referred to as a controller).

In 1967, the vast majority of the electronic data that pertained to individuals came directly from the individual’s actions. The individual would apply for a loan, register a deed, open an account, apply for a license, pay a bill or graduate from a school. All of these discrete actions would create a record that truly involved the individual. Within this setting, the actions were matched by a collection of data in which the individual participated. So, collection and origin were one in the same.

At the time, there were small observational data sets, but most were not computerized. Physicians created notes about their patients in paper-based medical charts, small merchants made notes about their best customers and early direct marketers noted similarities about their best customers. These mostly manual data sets—created without the involvement of the individual—were, for the most part, not significant enough to impact a governance model that was generally based on individual autonomy. The taxonomy will classify this category of data as observed.

As long as there has been data that pertains to an individual, there have been others that have looked for similarities in the data. Merchants have been classifying their customers based on common attributes for as long as there have been buyers and sellers. In 19th-century North America, merchants created co-ops to
share information about credit worthiness with classifications derived from shared data. The direct marketing industry began with the simple process of using transactional data to derive market segments based on look-a-likes. Furthermore, once analysts began looking for similarities, they began to conduct simple arithmetic calculations to enhance comparisons. For example, would ratios of mortgage debt to consumer debt demonstrate something interesting? The products of these simple calculations are data derived from underlying data. For hundreds of years, the insurance industry has looked at birth, death, occupations, location and lifestyle, and the industry derived actuarial tables that companies still use today for life insurance underwriting. While the classification builds on data that comes from interactions and transactions that involve the individual, the individual is not involved in the creation of the new data. The taxonomy will classify this data as derived.

An early application of computerized statistics against large personal data sets was the MDS bankruptcy score in the 1980s. The MDS score made use of computerized credit reports to predict the likelihood that an individual would go bankrupt over the next five years. The MDS credit score was not just a matching of attributes of those individuals that went bankrupt but, rather, a statistically based prediction that was validated using historic data. The resulting credit score is a piece of data based on the probability of a future event taking place that is linked to an individual. While the underlying data came from interactions with the individual, the individual had no involvement in the creation of the score. The classification for this data is inferred.

While derived and inferred data both create data from other data, they do so in a different manner with very different outputs. Derived data comes from simply looking at common attributes or doing simple calculations. Inferred data, however, is based on probability. One can score the likelihood of a good or bad event taking place. If the data says that six of ten individuals will behave in a certain fashion, you can be fairly sure six of ten will. That does not mean that you know which six will behave in a certain man-
The data is based on an absolute probability that some outcome will take place. Decision makers depend more strongly on inferred data rather than derived. In fact, while actuarial sciences were historically based on derived data, they are more likely today be based on probability-based inferences.

2.2. **Rapid expansion of data**

The rapid increase in computing power, decrease in communications costs and falling prices for storage all led to the expansion of data sets in the late 1980s and the 1990s. However, the most significant trigger for data expansion was the literal explosion of observational data that was sparked by the Internet in the 1990s. The Internet facilitated the collection of very granular information on how individuals behave. An observable action was no longer limited to registrations, purchases and filings, but now it also included the micro steps that leads up to those actions. The fact that an individual paused over a pixel becomes a recordable piece of data. Much of this observational data originates in a fashion not linked to a readily identifiable individual. However, it often links to an individual in a manner that lets the non-identified individual to be characterized. So, observational data leads to the creation of both derivations such as likely responder and inferences such as 90% chance the individual is a fraudster.

The 21st century has led to sensor technologies that make granular observation possible in the physical as well as virtual world. Every major shopping mall has CCTV cameras, and images can be and are transformed into data. Automobiles have sensors that read how drivers operate their vehicles. The combination of online and physical observations have facilitated the massive expansion of observational data. While this data begins with the actions of individuals, the individuals are not active partners in the origination itself.

Bruce McCabe published the research paper “The Future of Business Analytics” in 2007. In many ways, McCabe’s paper an-
nounced the beginning of the Big Data era. McCabe noted that unformatted data could now be used for analytics processes. This significantly expanded the amount of data that could be used for research, since data no longer had to be formatted in traditional fields. Diverse data sets could therefore be used to discover correlations that were less obvious in the past. Those correlations lead to predictions pertaining to individuals in almost any setting. Informatics is increasingly able to rank order individuals based on probability, which will lead to a rapid expansion of inferred data.

2.3. **Taxonomy based on origin**

In the prior section, the paper briefly described how the early work in privacy focused on the data that comes directly from the individual in a manner that involves the individual. It also discussed other forms of personal data that have a long history but only began to become impactful as technology facilitated automation. This section will begin with a table that lays out data classifications based on the manner in which the data originated.

Column 1 lists the major classifications based on how the data originates.

Column 2 contains sub-classifications which help to make the analysis more granular. For example, some levels of observation are anticipated, such as the active sub-classification, while others are oblivious to the individual, such as the passive sub-category.

Column 3 includes examples to assist the reader in relating the categories to the data world.

Column 4 provides a simple ranking based on how aware the typical individual might be based on the distance and manner of data origination. Legacy data governance is very dependent on individual awareness to both exercise consent as well as data review (access rights) and correction.
## Table 1: Data Categories Based on Origin

<table>
<thead>
<tr>
<th>Category</th>
<th>Sub-Category</th>
<th>Example</th>
<th>Level of Individual Awareness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provided</td>
<td>Initiated</td>
<td>Applications, Registrations, Public records such as licenses, Credit card purchases, Medical history as provided by individual</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Transac-</td>
<td>Bills paid, Inquiries responded to, Blood pressure or weight as recorded in clinical care setting, Public records such as court proceedings</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>-ional</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Posted</td>
<td>Speeches in public settings, Social network postings, Photo services, Video sites</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Engaged</td>
<td>Cookies on a website, Loyalty card, Enabled location sensors on personal devices, Fitness tracking using wearable device</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Not Antici-</td>
<td>Data from sensor technology on my Car, Time paused over a pixel on the screen of a tablet</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>-pated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Passive</td>
<td>Facial images from CCTV, Obscured web technologies, Wi-Fi readers in buildings that establish location</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Compu-</td>
<td>Credit ratios, Average purchase per visit, Risk of developing a disease based on a single genetic variation</td>
<td>Medium to Low</td>
</tr>
<tr>
<td></td>
<td>-tional</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Notational</td>
<td>Classification based on common attributes of buyers, Medical condition based on diagnostic tests</td>
<td>Medium to Low</td>
</tr>
<tr>
<td></td>
<td>Statistical</td>
<td>Credit score, Response score, Fraud scores, Life expectancy</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>Risk of developing a disease based multi-factor analysis, College success score based on multi-variable Big Data analysis at age 9</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Analytical</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4. Data category further description

2.4.1. Provided Data

Provided data originates via direct actions taken by the individual in which he or she is fully aware of actions that led to the data origination.

The taxonomy breaks the category into the three sub-categories of initiated, transactional and posted.

2.4.1.1. Initiated

Initiated data is the product of individuals taking an action that begins a relationship. These actions might include applying for a loan, registering to vote, taking out a license, providing past medical history to a new health care provider or registering on a website. The individual is aware of the action he or she is taking. While the individual does not always consider the implications, it would be intuitive to the individual that his or her actions would create data that pertains to him or her.

2.4.1.2. Transactional

Transactional data is created when an individual is involved in a transaction. Transactions may include buying a product with a credit card, paying a bill, responding to a question, having a blood pressure reading in a clinical care setting or taking a test. While the individual might not be thinking about the fact that he or she is creating a record, they understand the transaction must be recorded, records need to be updated and histories modified. The individual is an active participant in the origin of the data.

2.4.1.3. Posted

When individuals proactively express themselves, they are aware that they are creating expression that will be seen or heard
by others. In past years, the recorded data might be a newspaper or television story. The growth of social networks has actively increased the origination of data based on proactive speech. While the individual is not always aware of who might see or hear the expression, they are fully involved in its creation.

2.4.2. Observed Data

Observed data is simply what is observed and recorded. The emergence of the Internet as an interactive consumer medium has made it possible to observe and digitalize data in a more robust manner. On the Internet, one may observe where the individual came from, what he or she looks at, how often he or she look at it, and even the length of pauses. Facial recognition and the Internet of Things is making observation in a digital manner possible in the physical world. For the purposes of this analysis, I have three sub-categories based on the level of awareness by the individual.

2.4.2.1. Engaged

Engaged observed data includes data that originates from online cookies, loyalty cards and other instances in which the individual is made aware of the observation at some point in time. While the individual may forget that the data is being created, there is a general awareness that it is taking place. In some cases, the individual can object to or abort the creation. For example, a person may disable the Wi-Fi on their mobile device if they don’t want to be observed. In the case of wearable technologies, a person may simply stop wearing it. Regulation and industry practice have implications on which sub-classification a type of data might fit. For example, cookies are included in the engaged sub-category, because various regulations and industry codes have made transparency a growing norm.
2.4.2.2. Not Anticipated

Not anticipated data creation are instances in which individuals are aware that there are sensors but have little sense that the sensors are creating data that may pertain to the individual. For example, a person may be aware that there are sensors in the tires on the car and in the oil pan in the engine, but the person might not be aware that the manner in which he or she maintains the car is a data element that might pertain to them. This sub-classification would be appropriate for many of the applications related to the Internet of Things. Typical individuals would have limited awareness of this type of data.

2.4.2.3. Passive

The last sub-category is passively created observational data. An example is CCTV in public places when combined with facial recognition. It is also applicable to any situation in which it would be very difficult for individuals to be aware that they are being observed and data pertaining to the observation is being created.

2.4.3. Derived

Derived data is data that is simply derived in a mechanical fashion from other data and becomes a new data element related to the individual. There are two sub-categories of derived data.

2.4.3.1. Computational

Computationally derived data is the creation of new data element through an arithmetic process executed on existing numeric elements. For example, a lender might create a computational data by calculating the ratio of mortgage debt to total consumer debt, an online merchant might calculate average spend per visit or a merchant might calculate the percentage of returned items to items bought. A medical geneticist may calculate the suscepti-
bility of a person to a particular disease based on a mutation in a specific gene that controls the disease, such as cystic fibrosis. Each of the new computational products is a data element that might be used by an organization to better understand behaviour or make decisions pertaining to the individual. The individual would not typically be aware of the creation of the new data element.

2.4.3.2. *Notational*

Notionally derived data are new data elements created by classifying individuals as being part of a group based on common attributes shown by members of the group. Direct marketing segment codes used throughout Europe are an example. Segment codes are created by data brokers that look for common attributes of groups of buyers and link them together with a common designation.

2.4.4. Inferred

Inferred data is the product of a probability-based analytic process. This category name is the same as that used by the World Economic Forum. This category includes two sub-categories.

2.4.4.1. *Statistical*

Statistically inferred data is the product of characterization based on a statistical process. Examples include credit risk scores, most fraud scores, response scores, life expectancy and profitability scores. The individual is not typically involved in the development of these scores.

2.4.4.2. *Advanced Analytical*

Advanced analytical data are the product of advanced analytical processes such as those found in Big Data. These data elements
are typically the product of analysis on larger and more diverse data sets, and the elements are based on analysis that is more dependent on correlation rather causation. Early examples of such data elements are identity scores that predict the likelihood that an identity is real. While credit scores were dependent on looking at past credit failures and what correlated to and affected those failures, identity scores were based on anomalies in the manner in which identities were structured. This required a new type of analysis that had not been possible in the past.

In the medical field, Big Data is beginning to generate insights into the likelihood of future health outcomes. The individual would not be aware of the creation of these new data that are the product of the inferences that come from analysis.

2.5. Data begets data

Provided and observed data comes directly from the contributions of and the observations of individuals. Derived and inferred data are the products of processing other data. However, once created, derived and inferred data then become the feed stock for future data created by ongoing processing.

If one were trying to predict the growth patterns for data, one would postulate that growth in submitted data will be fairly flat. Individuals will only apply for so many loans, register at so many websites or pay so many bills. Growth in this category would probably be in the posted sub-category as individuals submit picture and postings.

Growth in observed data should continue to accelerate as a sensor-rich environment continues to be built out. Much of that growth will be in the unexpected and passive categories, so individual participation in its creation will be minimal.

Derived data, I believe, will have a flat growth curve as business processes become more robust and analysis becomes more sophisticated. In simple terms, derived data will be replaced by inferred data.
Inferred data will accelerate as more and more organizations, both public and private, increasingly take advantage of broader data sets, more computing power and better mathematical processes.

The bottom line is that data begets more data. That data is increasingly created at a distance from the individual and without the individual’s involvement. As noted in column 4 and described above, the data tends to be the product of more sophisticated processes, and its application has more positive implications for all parties involved. The application of the data also creates new risks that the individual is not in a position to mitigate via autonomy rights, such as consent.

3. POLICY ANALYSIS

The previous section established that origination is a useful lens to look at personal data. The fact is the further one moves from originating data via direct collection from the individual, the more challenging governance based on individual participation will be. Despite the fact that data beyond provided data have existed as long as there has been automated processing, and these types of data are accelerating, the proposed regulation makes no reference to data origination other than collection. This continues the precedent from the existing Directive and the 28 laws that implement it.

To fulfill the obligation that data be processed in a lawful, fair and transparent fashion, the recitals in the new regulation requires that data purposes should be defined at the time of collection, and the data not be excessive and only held as long as necessary. The regulation seems to anticipate a linear progression

---

2 For purposes of this analysis, the 25 January 2012 proposed regulation draft was used.
3 Recital (30).
from collection, to use and storage, and finally to destruction. When discussing further processing beyond the original specified purposes, the recital to the regulation referees to the purposes specified at collection. The recitals further states “where data are not collected from the data subject” in discussing purpose specification. This phrase seems to refer to collection that comes from a third party that has collected the data from the individual. There are no references to data that is created rather than collected.

When one reads the general revisions, one might begin with the definitions. There are definitions for many words but none for collection. Without a definition one is left with concept that a collection comes from the individual providing data or being observed. Collection does not seem to include derived and referred data.

The principles in article 5 includes one for collection. It says:

*Personal data must be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes*

One may read that section as requiring that the purpose of creating new data from the data collected must be done in a manner not incompatible with the purposes noted at collection. One could make the argument that any additional inferences should not be inconsistent with the original purposes. This is further supported by recital (40) which states that data used for research should be used in a manner compatible with the original purpose stated at collection.

Article 6 of the prosed regulation lists six legal basis for processing personal information. While one can see means for processing information to yield created data, and also to make use of that data in applications, the concept of created data is not explicit in the regulation.

The concept of transparency is built into requirements for purpose specification and individual participation is built into consent, access and correction, and the revoking consent. There is no clear path for transparency for created data other than purpose
specification notice that might reference data creation as a purpose.

Article 20 on measures based on profiling raises troubling issues related to inferred data. Depending on how profiling is defined, inferred data may well be the fruits of a profiling process. While the article seems to be clear that the issue is not the creation of the inference, but rather the application of an inference. In 2013, the Centre for Information Policy Leadership produced a paper that suggested a two-phase approach to advanced analytics that would divide the research phase (where inferences are developed) and the application phase (where inferences are used)\(^4\). The differences between research and application were recognized in the WP29 2013 paper on compatible uses. However, that work is not captured in the proposed regulation.

In sum, the proposed regulation was written without a recognition of the changed mix of data types that have become much more prevalent over the past 20 years.

### 4. FURTHER ANALYSIS

Will the proposed regulation capture the risks to individuals’ fundamental rights while facilitating the necessary movement of data for commerce, economic growth and free expression as the mix of data continues to trend to more observed and inferred data? Ultimately, that will depend on the flexibility built into the proposed regulation and the judgment shown by the new Data Protection Board (“DPB”) to be created by the regulation. The proposed legal instrument is a regulation rather than a directive. However, the proposed regulation still maintains one or more independent data protection authority in each state.

---

We cannot know at this time just how flexible the new regulation might be when enacted. While Parliament reached a common position prior to the writing of this paper, the Council had not. There are specific sections that will touch on the flexibility of the instrument. Those that I believe will be most relevant are those sections that will define the legal basis for processing and sections on profiling. Parliament was fairly restrictive and those areas, and the Council discussion is not complete. As stated earlier, consent is almost never an effective governance structure for governing inferred data, and I would argue is not overly effective in governing observed data.

I believe the DPB might be more effective in driving flexibility. To assure data protection harmonization, the proposed law creates the DPB. Unlike the current Working Party 29 the DPB will establish a common interpretation and implementation of the regulation. Common facts should lead to common decisions.

The current Working Party 29 acts as a policy development agent for European data protection by issuing opinions. The WP29 has created very useful guidance on accountability (2010)\(^5\) and compatible purposes (2013)\(^6\), anonymisation (2013)\(^7\), and legitimate interests (2014)\(^8\). While none of these papers directly addresses the new data classification, the papers indirectly address many of the issues raised by observed and inferred data. From the perspective of this analyst (not a lawyer), there is flexibility built into the various provisions to take the values related to collection and suggest guidance based on created data.

I believe that a DPB would continue to explore an update these key issues. While all opinions would have to be rooted in the law, I have seen evidence that authorities, when they understand the

\(^6\) Opinion 03/2013 on purpose limitation.
\(^7\) Opinion 05/2014 on “Anonymisation Techniques on the Web”.
\(^8\) Opinion 06/2014 on the “Notion of Legitimate Interests of the Data Controller under Article 7 of the Directive 95/46/EC”.  

facts in play, will protect individuals’ fundamental rights while also looking to practical interests.

There is one final note. Markets driven by fair information require market knowledge. There is general agreement that purpose specification notices are not sufficient to provide individuals and communities of individuals (often referred to as the “crowd”) with the knowledge necessary to assure individual participation, a basic fair information value. I do not believe the proposed regulation, or other draft laws I have seen, address this issue. Governance of these new data sets will not be completed until that issue is first explored and then address. Better transparency is a challenge for the entire privacy community including industry, civil society and academics.

For the proposed regulation to be effective, those that implement and interpret must understand the challenges related to created data.