THE Information Accountability Foundation
COLLABORATIVE POLICY INNOVATION

"The Origins of Personal Data and its Implications for Governance"

By Martin Abrams

The Information Accountability Foundation

Executive Summary

- Legacy privacy governance regimes are based on a presumption that data is primarily being collected from the individual with some level of their awareness.
- Increasingly data is not collected directly from the individual but, rather, at a distance without the individual's awareness of its origination and subsequent uses.
- To understand the implication, this paper proposes a taxonomy based on the manner in which data originates. The data categories include:
    - Provided
    - Observed
    - Derived
    - Inferred

Introduction and Purpose

Data constitutes the life blood 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.

The OECD documented the expansion of data and its uses in "The Evolving Privacy Landscape: 30 Years After the OEC Privacy Guidelines." The 2011 paper was published to inform the experts to make recommendations on further development of the very successful OECD Privacy Guidelines. The paper makes the case that communications and computing technologies have made more things possible, that more data flows globally, the Internet and sensors increase the amount of data, and business processes have changed to take advantage of the rapid expansion in data.

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. To get governance right, we must understand where data comes from, how it is created, and how aware and involved the individual is in its creation.

The purpose of this paper is to create a taxonomy of data based on how it first originates and tracks the policy issues that arise with new data types.[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, and 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 would be 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, 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 one exception was investigative consumer reports, in which the observations of individuals were collected as part of a report for purposes such as employment.

[1] Origin is not the only lens one might use to classify data. The OECD Digital Economy Papers No. 220, "Exploring the Economics of Personal Data," contains a taxonomy of data based on the concept of data collection borrowed from the World Economic Forum. The taxonomy looks at the data from a collection perspective related to a data lifecycle. The OECD paper also references Bruce Schneier's "Taxonomy of Social Networking Data" that was revised in Schneier's blog on 10 August, 2010. Schneier's taxonomy does an excellent job of cataloging data from the perspective of social networking. The OECD paper also references classifications based on the nature of the relationship of the individual to the collector.

In the United States, those reports were and still are governed by the Federal Fair Credit Reporting Act. 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 has 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 product of these simple calculations are data derived from underlying data. 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.

The first application of statistics against larges 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.

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 observationed 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, filings but 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 and are transformed into data. Automobiles have sensors that read how the vehicle is operated. The combination of online and physical observation 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 announced the beginning of 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 where 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.

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 is 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, 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.

Table 1: Data Categories Based on Origin

Category	Sub-Category	Example	Level of Individual Awareness
Provided	Initiated	○ Applications ○ Registrations ○ Public records ○ Filings ○ Licenses ○ Credit card purchases	High
	Transactional	○ Bills paid ○ Inquiries responded to ○ Public records ○ Health ○ Schools ○ Courts ○ Surveys	High
	Posted	○ Speeches in public settings ○ Social network postings ○ Photo services ○ Video sites	High
Observed	Engaged	○ Cookies on a website ○ Loyalty card ○ Enabled location sensors on personal devices	Medium
	Not Anticipated	○ Data from sensor technology on my Car ○ Time paused over a pixel on the screen of a tablet	Low
	Passive	○ Facial images from CCTV ○ Obscured web technologies ○ Wi-Fi readers in buildings that establish location	Low
Derived	Computational	○ Credit ratios ○ Average purchase per visit	Medium to Low
	Notational	○ Classification based on common attributes of buyers	Medium to Low
Inferred	Statistical	○ Credit score ○ Response score ○ Fraud scores	Low
	Advanced Analytical	○ Risk of developing a disease based multi-factor analysis ○ College success score based on multi-variable big data analysis at age 9	Low

Data Category Further Description

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 three sub-categories, initiated, transactional, and posted.

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, or registering on a website. The individual is aware of the action he or she is taking. While the individual doesn't 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.

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, 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.

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.

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.

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. Regulation and industry practice have implications on which sub-classification a type of data might fit. For example, cookies are included in engaged because various regulations and industry codes have made transparency a growing norm.

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.

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.

Derived

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

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. Each of the new computational products is a data element that might be used by an organization to better understand behavior or make decisions pertaining to the individual. The individual would not typically be aware of the creation of the new data element.

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. For example, a marketer might notice its customers have six common attributes and look for the same attributes in a group of potential customers.

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.

Statistical

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

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 impacted 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.

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. 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.

Key Policy Questions

In 2013, the OECD updated its privacy guidelines, first adopted in 1980. Revising the guidelines, the OECD added additional guidance on accountability. The wording of the guiding principles remained fundamentally the same as adopted in 1980 and links governance to collection. This creates challenges for applying the principles to the manner in which data originates today. This section will briefly look at each of the principles and raise possible questions for the OECD to consider.

Collection Limitation

As noted in this paper, data increasingly is created not collected. Does the OECD focus on collection make the principle less useful? If one looks beyond the principle's structure, the issues raised by the principles, lawfulness and fairness, are even more relevant in the current data rich world. The principle also acknowledges that not all data originates in a manner where consent and knowledge are applicable. The principle also points to the need for greater individual awareness. However, the structure, focused on collection, raises questions on how those underlying issues of lawfulness and fairness might be applied to the current data classes.

Data Quality

Data quality is very relevant to the current discussion. No matter how data originates it should be appropriate for its uses. Future OECD guidance pursuant to Big Data may want to explore the governance challenges related to data quality.

Purpose Specification

Purpose specification has had two objectives over the past 34 years. The first is to provide transparency to the individual about how data will be used. The second is to provide discipline to the data user about future scope of use. With data originating at a distance and without the explicit knowledge of the individual, purpose specification is less functional as a transparency tool. The second discipline, future guidance for application seems very relevant. So a question arises on how to achieve both objectives, transparency and discipline with the guidelines if not the principles.

Use Limitation

This principle raises the same issues as the previous one. Big data processes pull data into applications to both discover trends and then build applications based on the newly identified trends. Origination of new data is sometimes the byproduct of those processes. Previous work by the Big Data Project at the Centre for Information Policy Leadership discussed governance related to discovery versus application. Future OECD guidance related to privacy and Big Data may want to suggest the manner in which this principle might be applied.

Security Safeguards

Data no matter how it originates should be secure proportional to the risks associated with the data. Future OECD guidance related to security safeguards might want to reiterate the importance of security safeguards as it relates to data that originates as part of analytic processes.

Openness

Openness to the creation and use of data is increasingly important. A key question is how that might be achieved. Transparency at point of collection is relatively easy compared to transparency pertaining to data processes that are not readily apparent. The author believes additional time and resources should be dedicated to increased transparency.

Individual Participation

The author believes individual participation is also very relevant to data originating at a distance from the individual. In many ways, the issues linked to individual participation are linked to the openness principle. The question isn't whether individuals should have the right to see data and challenge underlying data but how the mechanisms to achieve the objectives of this principle might be designed.

Accountability

Accountability is the key principle in assuring governance when data originates at a distance from the individual. The additional guidance contained in the 2013 revisions are most useful. However, there is room for even more commentary on how to be accountable. Some of the commentary has been developed by privacy enforcement agencies in Canada and more recently Hong Kong.

In summary, the growing proportion of observed and inferred data challenges the concept that the nexus for governance is collection and the assumption that awareness goes naturally with collection. The OECD might want to consider additional work to tie the objectives of the privacy principles to data that originates at a distance from the individual without the individual's participation and awareness.

About the Author

Martin Abrams is Executive Director of The Information Accountability Foundation. 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.

Comments

The foundation considers data classification as a work in progress. The concepts in the paper need to be tested, and the policy implications debated. Please send your comments to mabrams@informationaccountability.org.