Big Privacy:

Bridging Big Data and the Personal Data Ecosystem Through *Privacy by Design*

December 2013

Ann Cavoukian, Ph.D.
Information and Privacy Commissioner
Ontario, Canada

Drummond Reed
Co-Founder and CEO
Respect Network

Acknowledgements

The co-authors would like to gratefully acknowledge the contributions of the following individuals, whose efforts were invaluable in the drafting of this paper: Dan Blum, Principal Consultant & Chief Security and Privacy Architect, Respect Network; Michelle Chibba, Director of Policy and Special Projects, IPC; David Weinkauf, Policy & Information Technology Officer, IPC; and Gary Rowe, Executive Chairman, Respect Network.

Information and Privacy Commissioner
Ontario, Canada

2 Bloor Street East
Suite 1400
Toronto, Ontario
M4W 1A8
Canada

416-326-3333
1-800-387-0073
Fax: 416-325-9195
TTY (Teletypewriter): 416-325-7539
Website: www.ipc.on.ca
Privacy by Design: www.privacybydesign.ca

Big Privacy:

Bridging Big Data and the Personal Data Ecosystem Through *Privacy by Design*

TABLE OF CONTENTS

1. Introduction

Recent technological and business developments have given rise to a new understanding of personal information. It is now being compared to currency and energy[1]—often being referred to as "the new oil."[2] It is an economic asset generated by the identities and behaviors of individuals and their technological surrogates. These metaphors, which express its increasing economic value to organizations, ring especially true in the case of Big Data. Indeed, Big Data derives economic value from its use of personal information to such an extent that if personal information is considered to be "the new oil," then Big Data is the *machinery* that runs on it.

However, like our current dependence on fossil fuels, Big Data's current use of personal information is unsustainable, increasingly resulting in "pollution" via privacy infringement. At the moment, individuals have little, if any, control over their information's use and disclosure in Big Data analytics. In addition to a host of privacy concerns, this lack of informational self-determination gives rise to an uneven exchange of the economic value. While the owners of Big Data algorithms profit from their use and disclosure of personal information, the individuals the personal information relates to do not—at least not directly. If not properly addressed, the privacy and economic concerns raised by Big Data threaten to decrease individuals' willingness to share their personal information[3]—in effect, cutting off the flow of the "oil" on which the analytic "machinery" of Big Data runs.

In order to make the interactions between Big Data and individuals more sustainable—in other words, to effectively transform this "new oil" into a renewable resource—the concept of the personal data ecosystem (PDE) has been proposed.[4] The PDE is the emerging landscape of companies and organizations that believe individuals should be in control of their personal information and directly benefit from its use, making available a growing number of tools and technologies to enable such control.

We have discussed elsewhere how the *Privacy by Design* framework can be applied individually to the PDE and to Big Data systems in order to achieve positive-sum, "win-win" solutions in which individuals maintain control over their personal information flows, without diminishing system functionality.[5] The time has come, however, to show how *Privacy by Design* can be applied at a larger scale to individuals' online interactions by bringing together the user-centric architecture of the PDE with the analytic power of Big Data, resulting in a much larger "win-

1 C. Moiso, R. Minerva. (2012). "Towards a User-Centric Personal Data Ecosystem," *Paper presented at the 16th International Conference on Intelligence in Next Generation Networks,* Berlin, Germany.

2 M. Kuneva. (March 31, 2009). *Keynote Speech: Roundtable on Online Data Collection, Targeting and Profiling*, Brussels.

3 See A. Mantelero. (2013). "Competitive value of data protection: the impact of data protection regulation on online behaviour." *International Data Privacy Law.*

4 See World Economic Forum. (2011). *Personal Data: The Emergence of a New Asset Class.* Retrieved from http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf.

5 See A. Cavoukian. (2012). "*Privacy by Design* and the Emerging Personal Data Ecosystem." Retrieved from http://www.ipc.on.ca/images/Resources/pbd-pde.pdf; A. Cavoukian. (2013). "Personal Data Ecosystem (PDE) – A *Privacy by Design* Approach to an Individual's Pursuit of Radical Control." *Digital Enlightenment*; A. Cavoukian, J. Jonas. (2012). "*Privacy by Design* in the Age of Big Data." Retrieved from http://www.ipc.on.ca/images/Resources/pbd-big_data.pdf.

win" scenario, where privacy and Big Data may coexist in tandem. This new result is what we call *Big Privacy*.

In this paper we will:

- Summarize the Big Data privacy challenges and the growing need to restore trust;
- Provide a definition of *Big Privacy* that parallels the definition of Big Data;
- Define the seven major architectural elements of *Big Privacy*:
  - Personal clouds
  - Semantic data interchange
  - Trust frameworks
  - Identity and data portability
  - Data-by-reference (and subscription)
  - Accountable pseudonyms
  - Contractual data anonymization
- Offer an exemplar showing how these elements are being implemented by the Respect Network using the OASIS XDI semantic data interchange protocol;
- Summarize how the seven architectural elements of *Big Privacy* apply the 7 Foundational Principles of *Privacy by Design* to the unique privacy challenges of Big Data.

2. Big Data, Privacy Challenges, and the Need to Restore Trust

It is worth considering how a term that no one had heard of a few years ago is now so popular it appears on airport billboards around the world: **Big Data**. What does it actually mean?

Definition of Big Data

The first sentence of the Wikipedia article captures the essential concept:

> ***Big data*** *is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.*[6]

6 "Big data." (n.d.). In *Wikipedia*. Retrieved October 8, 2013, from http://en.wikipedia.org/wiki/Big_data

What's missing from this capsule definition is the context that Big Data only emerged as a concept after the convergence of a host of new technologies capable of processing these very large data sets. These technologies made it possible to capture value that could not be realized before. And it is a surprisingly long list of technologies—see the list highlighted below from this excerpt of the Wikipedia definition:

> *Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report suggests suitable technologies include A/B testing, crowdsourcing, data fusion and integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis, and visualization.*
>
> *Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data-mining grids, distributed file systems, distributed databases, cloud-based infrastructure (applications, storage, and computing resources), and the Internet.*[7]

The Privacy Challenges of the Big Data Life Cycle

Any technology or operational practice that operates upon very large data sets of personally identifiable information (PII) invites heightened privacy concerns. For example, the clear trend in Big Data towards increased collection, storage, and linkage of personal data sets raises privacy concerns relating to theft, institutional misuse, and re-identification, as well as unauthorized access, redistribution, and repurposing of personal information. In order to prevent these concerns from becoming real harms, effective policy and technological measures are required on the part of organizations that use Big Data analytics, as well as for individuals to whom the data relates. The IPC has produced two papers on how to address the privacy challenges of Big Data in different contexts:

- "*Privacy by Design* in the Age of Big Data,"[8] which outlines how an advanced Big Data sensemaking technology was engineered, from the ground up, with privacy-enhancing features; and
- "Introducing Privacy-Protective Surveillance: Achieving Privacy *and* Effective Counter-Terrorism,"[9] which advances a positive-sum alternative to current counter-terrorism surveillance systems and shows how its application of specific "feature/event-based" data analysis through the use of intelligent virtual agents, homomorphic encryption, and probabilistic graphical models can be more effective than non-targeted and broad "pattern-based" data mining, the application of which is questionable in highly sensitive areas such as counter-terrorism.

7 Ibid.

8 A. Cavoukian, J. Jonas, (2012). "*Privacy by Design* in the Age of Big Data." Retrieved from http://www.ipc.on.ca/images/Resources/pbd-big_data.pdf.

9 A. Cavoukian, K. El Emam. (2013). "Introducing Privacy-Protective Surveillance: Achieving Privacy and Effective Counter-Terrorism." Retrieved from http://www.ipc.on.ca/site_documents/pps.pdf.

Generally speaking, the privacy challenges of Big Data occur across all three phases of its iterative life cycle: (1) data harvesting, (2) data mining, and (3) application—which we refer to as the **Big Data life cycle** (BDLC).[10]

1. **Data Harvesting.** As its name implies, Big Data requires a sufficiently large data set in order for its analytics to generate algorithms that capture new and unanticipated connections among the data analyzed. The *data harvesting phase* of the BDLC is where such a sufficiently large data set is amassed from various sources including data actively provided by users, passively generated by their devices, or harvested online.

2. **Data Mining.** Once a sufficiently large data set has been amassed, it may be analyzed in the hopes of discovering connections among the data that will lead to an advancement in knowledge. This process of discovery constitutes the *data mining phase* of the BDLC. The result of the data mining phase is a correlation that generalizes the relationships between the discovered connections in the form of an algorithm, which produces an output that specifies the degree to which a particular data set exhibits characteristics of the knowledge advanced by the correlation.

3. **Application.** A Big Data algorithm by itself is simply the generalized form of a discovered piece of knowledge. In order for its knowledge to be useful, the algorithm must be applied to the "real world" where something is at stake. The *application phase* of the BDLC is where an algorithm generated in the data mining phase is then applied to individual data sets in order to produce insights about a real-world situation. Depending on the data set input into the algorithm, the insights produced may disclose personal information about the individuals associated with the data. The following quote from Foster Provost and Tom Fawcett's book, *Data Science for Business*, illustrates some useful applications of Big Data algorithms:

   "For example, one fundamental concept [of Big Data] is that of determining the similarity of two entities described by data. This ability forms the basis for various specific tasks. It may be used directly to *find* customers similar to a given customer. It forms the core of several *prediction* algorithms that estimate a target value such as the expected resource usage of a client or the probability of a customer to respond to an offer. It is also the basis for *clustering* techniques, which group entities by their shared features without a focused objective. Similarity forms the basis of *information retrieval*, in which documents or webpages relevant to a search query are retrieved. Finally, it underlies several common algorithms for *recommendation*."[11]

10 The phases presented here are used for the purpose of illustrating the privacy challenges associate with Big Data. As such, they represent a simplification of its overall process. For more complex and technical discussions of Big Data and its life cycle, see U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. (Fall 1996). "From Data Mining to Knowledge Discovery in Databases." *AI Magazine*. Retrieved from http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf; and Centre for Information Policy Leadership, Hunton & Williams LLP. (February 2013). "Big Data and Analytics: Seeking Foundations for Effective Privacy Guidance." http://www.hunton.com/files/Uploads/Documents/News_files/Big_Data_and_Analytics_February_2013.pdf.

11 F. Provost, T. Fawcett. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, p. xii.

Each phase of the BDLC brings with it unique privacy challenges. For example, the data harvesting phase (phase 1) attempts to amass a sufficiently large data set for data mining. Often times this amounts to a "grab everything in sight" or "collect first, ask questions later" approach to data acquisition. Such approaches, however, fly in the face of the collection limitation privacy principle, according to which there should be *limits* placed on the collection of personal information.

In addition, during the data mining phase (phase 2), a new and unanticipated connection is made. Precisely in making such a connection, however, this phase goes beyond the original purpose for which the data was collected. Thus, if the data set contains personal information, such a production of knowledge would conflict with purpose specificity, which states that personal information should not be used for purposes other than those specified at the time of collection or those which are not incompatible with the original purpose. Otherwise consent of the individual is required.

Finally, in applying an algorithm to an individual data set, the application phase (phase 3) will produce insights about a real-world situation, which may disclose personal information about the individuals associated with the data. If the individual the data set pertains to has not consented to such disclosure, however, this practice would conflict with use limitation, which confines the uses of one's personal information to those originally specified at the time of collection, unless the individual has consented to such other uses. In the case of Big Data, such a consent option is rarely offered to individuals, let alone given by them. Indeed, the fact that the use of Big Data algorithms is not known until their actual discovery makes it difficult, if not impossible, to obtain consent from individuals for secondary uses.

These challenges show how privacy cannot be assured solely by compliance with regulatory frameworks, which tend to lag behind recent technological developments such as Big Data. While standard privacy principles are an essential tool for protecting the privacy of individuals, as a standalone measure, they are nonetheless insufficient. In order to address the challenges of new technologies such as Big Data, something more dynamic is needed—something which is able to stay ahead of the ever-advancing "curve" of technology. We believe this "something" is *Privacy by Design*. When applied to the case of Big Data, this means that in addition to a regulatory framework, what is needed is an *ecosystem* of technologies and policies that affords individuals a far greater, indeed "radical" degree of control over their personal information, from which they may derive considerable benefits.

Restoring Trust with the Personal Data Ecosystem

Since one of the 7 Foundational Principles of *Privacy by Design* is protection of personal data across its entire life cycle, with Big Data, we believe this can only be accomplished through privacy solutions that are equally encompassing. This is where the personal data ecosystem (PDE) comes in. As mentioned in the Introduction, the PDE is the set of companies, organizations, and policymakers who believe that individuals should be in control of their personal information and who are employing new tools, technologies, and policies to empower them.

The emergence of the PDE at this juncture is no coincidence. Big Data and other powerful new analytical systems can only exist (comfortably) in their current form so long as people are largely unaware of the scope and extent to which these systems gather, use, and trade in PII. But the "Summer of Snowden" has changed all that. Like a dam breaking, these revelations about the U.S. and other government's Big Data gathering and analytic capabilities dramatically raised awareness of how seriously they impinge on individual privacy, not only in the United States, but worldwide.

So if privacy infringement is the negative externality that Big Data frequently ignores, the PDE is the emerging positive externality that can turn the combination into a positive-sum outcome where both data subjects and Big Data users benefit. The greater the challenges to privacy and self-determination from Big Data and other quarters, the stronger the PDE will rise up to meet it, which brings us to the definition of *Big Privacy*.

3. A Definition of *Big Privacy*

Our definition of *Big Privacy* directly parallels that of Big Data:

> ***Big Privacy*** *is* Privacy by Design *writ large, i.e., it is the application of the 7 principles of* Privacy by Design*, not only to individual organizations, applications, or contexts, but to* ***entire networks, value chains, and ecosystems,*** *especially those that produce and use Big Data. The goal of Big Privacy is the* ***systemic protection*** *of personal data and* ***radical personal control*** *over how it is collected and used. Radical control is an embodiment of "informational self-determination"—the right enshrined in the German Constitution relating to the individual's ability to determine the fate of one's information.*[12] *This means that it must be possible to assure whole populations that their privacy is being respected because the network, value chain, and/or ecosystem producing and processing Big Data has implemented* Privacy by Design *at a system-wide level, enabling individuals who consent to the use of their personal information to reap a proportion of the benefits.*

It is important to note that in the context of our growing dependence on information and communications technologies (ICTs), privacy does not equal secrecy of personal data, it equates to **individual control** of one's data. This does not mean that organizations cannot use, or have no meaningful relationship to, individuals' personal data; only that organizations' processing of personal data must have user-centric controls in place such as informed consent, purpose specificity, use limitation, etc. Privacy is not about keeping information secret (hiding information), but rather about having a right to "informational self-determination," which may or may not result in an individual sharing his/her personal information. However, when one's personal data is shared by consent, and used to generate an economic benefit by Big Data users, a portion of that benefit will accrue to the user. In this respect, privacy is highly contextual and will manifest itself differently, depending on the preferences of the individual and the different technologies, cultures,

12 The term "informational self-determination" was first used in a German constitutional ruling concerning personal information collected during Germany's 1983 census.

and situations to which it is applied. Within the limits of applicable laws, in a particular context, individuals must be able to choose whether or not to share:

- Everything about themselves
- Nothing about themselves
- Anything in-between.

Further, to make privacy fully operational means that individuals **must be able to change their mind at any time** about what is being shared, and with whom. This clarification—that privacy is all about control—is critical to this discussion: Big Data does not necessarily need to get *smaller* to accommodate the demands of the PDE, but it does need to get ***smarter***. In addition to harvesting (or not harvesting) personal data, it must also be able to access individual preferences about the data's use—or render the point moot using de-identification tools and other technologies.

There is a second important parallel between the definition of *Big Privacy* and Big Data. Just as Big Data emerged through the convergence of a set of powerful new data processing technologies that were able to handle much larger and more distributed data sets than ever before, *Big Privacy* is emerging through the convergence of new PDE technologies that give individuals greater visibility and more control over their personal data than ever before. In fact, *some of these are the very same technologies that help to power Big Data.*

These technologies must solve complex, highly contextual problems. Not all personal data is created or wholly owned by individuals. Individuals should have privacy rights to control access to data such as government identifiers, bank account balances, medical test results, and purchasing history, but organizations that generated them may also have a strong interest in relation to that data. Fortunately, for the purposes of analytics, many of the same forces that have released the Big Data "genie" can, in fact, be used—not to put it back in the bottle—but to harness its power to help individuals better control, express, benefit from, and enforce their personal privacy preferences in relation to organizations. In this way, *Big Privacy* is empowered to be a positive-sum, "win-win" approach to privacy *and* Big Data (never underestimate the power of "and"). By using some of the same technologies as Big Data and by embedding them into the design and architecture of the PDE, *Big Privacy* allows for the legitimate interests and objectives of privacy and Big Data to coexist, with both values being achieved in tandem, instead of being positioned as opposing forces.[13]

4. The Seven Architectural Elements of *Big Privacy*

In this section we will describe the seven major new tools, technologies, and policy innovations that are converging to enable *Big Privacy*. We will explain where each of these is in its evolutionary development and how they mutually complement and reinforce one another.

13 See Section 6 for a further discussion of the positive-sum outcomes of *Big Privacy*.

1. Personal Clouds

In March 2012, Gartner announced that:

> *The reign of the personal computer as the sole corporate access device is coming to a close, and by 2014, the personal cloud will replace the personal computer at the center of users' digital lives.*[14]

The announcement explained that the personal cloud—a virtual compute space in the cloud controlled entirely by an individual and available to all of that person's devices—is becoming as inevitable as the personal computer was in the 1980's. The report goes on to cite five major trends supporting this evolutionary leap. For example:

> *The advent of the cloud for servicing individual users opens a whole new level of opportunity. Every user can now have a scalable and nearly infinite set of resources available for whatever they need to do. Users' digital activities are far more self-directed than ever before. Users demand to make their own choices about applications, services and content, selecting from a nearly limitless collection on the Internet. This encourages a culture of self-service that users expect in all aspects of their digital experience. Users can now store their virtual workspace or digital personality online.*[15]

Gartner is not alone in the industry with this perspective. Doc Searls, founder of ProjectVRM at the Harvard Berkman Center for Internet and Society[16] and author of *The Intention Economy: When Customers Take Charge*[17] summarizes the inevitability of personal clouds this way:

1. **First computers got personal.** This was the PC revolution of the 1980's.
2. **Then the Internet got personal.** This was the Web revolution of the 1990's.
3. **Then phones got personal.** This was the smartphone revolution of the last decade.
4. **Now clouds are getting personal.** And this revolution will happen faster than any of the previous steps because *it's all in the cloud.*

Nowhere is the emergence of personal clouds more relevant than the assertion of personal privacy on the Internet. The reason is that personal clouds are the first tool that gives individuals the ability to communicate as peers with companies, organizations, and governments. During a talk at Privacy/Identity/Innovation 2013, entitled *How Personal Clouds Can Bring about a Sea-Change in Internet Privacy*,[18] Respect Network founder Drummond Reed put it this way:

14 https://www.gartner.com/newsroom/id/1947315

15 Ibid.

16 http://blogs.law.harvard.edu/vrm/

17 http://www.amazon.com/Intention-Economy-When-Customers-Charge/dp/1422158527/

18 http://igniteshow.com/videos/how-personal-clouds-can-bring-about-sea-change-internet-privacy/

Personal clouds fix the power asymmetry that exists today between individuals and businesses on the Internet. Companies have big iron, Big Data, big software, and big staffs. Individuals have none of that. But with personal clouds, individuals can finally tap the power of the cloud for themselves, and begin to exchange data with companies as a first-class peer.

Personal cloud technology has itself been evolving rapidly. The earliest forms, called **personal data stores**, **lockers**, or **vaults**, have been on the market for several years now from vendors like Personal.com, Mydex, and OwnYourInfo.[19] More recently, larger vendors have been moving into the space, spurred by the success of Dropbox, Google Drive, and Apple's iCloud. And open source alternatives such as OwnCloud, remoteStorage, Cloud OS, and XDI2 are also advancing quickly. So Gartner's prediction that "the personal cloud will replace the personal computer at the center of user's digital lives by 2014" looks very close to the mark.

2. Semantic Data Interchange

Personal computers were a huge productivity boon when they reached the mass market in the 1980's. However, that paled by comparison to what happened after they became networked in the 1990's. Nowadays, a PC that is not connected to the Internet can no longer do many of the tasks that people perform every day, e.g., email, Web browsing, social networking (to test that, just try turning off the network on your computer for a day). And of course, the very newest personal computer—the smartphone—makes connectivity the very foundation of the platform.

The same will be true of personal clouds. As useful as they are for enabling an individual to easily share data across his/her own devices and applications, the real value will come from being able to share data securely and privately *between* personal clouds—and between personal clouds and business clouds.

However, just as it took a wide-area networking protocol like TCP/IP to enable the Internet, a wide-area email protocol like SMTP to enable global email, and a wide-area hypertext protocol like HTTP to enable the Web, it will take a wide-area data sharing protocol to enable private, secure cloud-to-cloud data sharing. This new type of protocol is called **semantic data interchange** because, rather than exchanging data in low-level binary formats (e.g., TCP/IP), or text-based markup languages (e.g., HMTL or XML over HTTP), it is exchanged in semantic graph formats using the Resource Description Framework (RDF), JavaScript Object Notation Linked Data (JSON-LD), or XRI Data Interchange (XDI) standards.

The first advantage of these semantic graph formats is that they enable a rich variety of data to be much more easily understood and integrated across widely disparate systems—an essential feature for the interoperability of the personal data ecosystem. However, from a privacy perspective, there is a second more important advantage: semantic data interchange can enable **machine-readable descriptions of the privacy preferences and permissions** associated with

19 For a detailed description of the offering from Personal.com, see A. Cavoukian. (2012). "*Privacy by Design* and the Emerging Personal Data Ecosystem." Retrieved from http://www.ipc.on.ca/images/Resources/pbd-pde.pdf.

shared data to travel with the data and be enforced by any system that comes in contact with it.

This feature, known as **link contracts** in XDI,[20] is a potential privacy game-changer. For the first time, the parties to a data sharing relationship will have a standard way to share portable, machine-readable policies governing the privacy, security, usage, and reuse of that data. Further, by adding semantic data interchange to personal clouds, we now have a way to create globally-scalable peer-to-peer (P2P) networks where every single individual can exercise a primary benefit of *Big Privacy*: *ongoing, sustainable control* of the personal data he/she has shared with any other personal or business cloud on the network.

While semantic data interchange is a significant step up from past methods of sharing context-free data, by itself it does not provide human-like contextual processing of data. But continuing advances in artificial intelligence, where autonomous agents will be able to contextually process data more like humans, will give rise to the next wave: a combination of mobile and cloud-based autonomous agents that can cooperate to help people make even more relevant and effective data sharing decisions.[21]

3. Trust Frameworks

What personal clouds and semantic data interchange are to the technology layer of the PDE, **trust frameworks** are to the legal and policy level. A trust framework is a new mechanism for achieving large-scale trust online that consists of two parts:

1. **The "tools"** are the technical standards and protocols that must be implemented by the members of a trust community to achieve on-the-wire interoperability;

2. **The "rules"** are the business, legal, or operational policies that must be followed in order to achieve the level(s) of security, privacy, and other trust assurances that the participants in the trust framework desire.

The trust framework itself is an online document that publishes the tools and rules together with an assessment and enforcement infrastructure that operationalizes them. In most cases the document serves as a new type of contract that legally binds the members of the trust community to the policies. In this fashion, a trust framework operates on a much higher level than site-specific privacy policies or Terms of Service (TOS).

Trust frameworks have only become established as major new tools for online trust after the formation of the Open Identity Exchange (OIX), the first international non-profit organization for the development of digital trust frameworks, in January

20 https://wiki.oasis-open.org/xdi/LinkContractPattern

21 See G. Tomko, D. Borrett, H. Kwan, G. Steffan. (August 2010). "SmartData: Make the data 'think' for itself." In *Identity in the Information Society*. Volume 3, Issue 2, pp. 343–362. Retrieved from http://link.springer.com/article/10.1007%2Fs12394-010-0047-x; I. Harvey, A. Cavoukian, G. Tomko, D. Borrett, H. Kwan, D. Hatzinakos (eds.). (2013). *SmartData: Privacy Meets Evolutionary Robotics*. Springer.

2010.[22] Other industry and academic efforts in this area include the Kantara Initiative[23] and the InCommon Federation.[24] The following chart summarizes the progress of new trust frameworks over the past three years.

Type	Trust Framework	Home	Status
Government-Initiated	U.S. FICAM	OIX	Live (2010)
	U.K. Identity Assurance	OIX	In progress
	U.S. NSTIC	OIX	In progress
Industry-Initiated	Kantara Identity Assurance	Kantara	Live (2010)
	Telecom Data Verification	OIX	In progress
User-Centric	Respect Trust Framework	OIX	Live (2011)
	Mydex Trust Framework	OIX	In progress

While all trust frameworks are welcome additions to the Internet trust landscape, from the perspective of the PDE, the most important ones are **user-centric trust frameworks**. Examples include the Mydex Trust Framework[25] and the Respect Trust Framework (discussed in greater detail in the next section). Both are designed expressly to enable the users of personal clouds to share data with each other, businesses, government agencies, and other relying parties with strong assurance that its security and privacy will be respected.

There is particularly strong synergy between personal clouds, semantic data interchange, and user-centric trust frameworks because personal data shared using controls such as XDI link contracts can bind the shared data to the policies of the trust framework. This means that users in the PDE will rarely need to be concerned about the privacy or security policies of a specific site because they can place their confidence in the trust framework(s) the site has agreed to. This is a perfect example of the *systemic, end-to-end protection* that is a hallmark of *Big Privacy*.

To operationalize a trust framework, the individuals, businesses, communities, and service providers participating in it must establish a trust network for which the trust framework specifies the necessary business, legal, and technical agreements. In theory, a trust network can operate on a purely peer-to-peer fashion with no centrally provided services, or it can be supported by specially-chartered service intermediaries. These intermediaries may provide registration,

22 http://www.openidentityexchange.org/

23 http://kantarainitiative.org/

24 http://www.incommonfederation.org/

25 http://openidentityexchange.org/trust-frameworks/mydex-trust-framework

discovery, reputation, billing, and other shared services. They should also have neutrality and privacy genes baked into their organizational DNA, much like the service providers that run today's telephone number or DNS registries.

Looking to the future and development of intelligent agents, the goal will be to embed the "tools" and "rules" of a trust framework within an intelligent agent ("SmartData") which becomes the personal surrogate of the data subject. This is another reason why AI developments must be folded into the PDE.[26]

4. Identity and Data Portability

The global banking system as we know it today would never have evolved if it did not give people the freedom to choose their bank and to move their money between banks at any time they wish. The same was *not* true of telephone companies: for many years telephone numbers were not portable. However, with the ascendance of mobile phones, the need for telephone number portability finally could not be repressed. In the United States, Local Number Portability (LNP) was mandated by the FCC in 1996.[27]

Note that telephone number portability still did not mean data portability. For example, even if you could port your telephone number to another carrier, you could not port other data such as your customer records and usage history. That limitation will simply never work with personal clouds. No individual will entrust storing his/her vital personal data and credentials with a Cloud Service Provider (CSP) who does not guarantee the personal cloud owner the right to move 100 percent of that data to a new CSP—or to self-host it—at any time.

This guarantee must also extend to the identifier(s) used by an individual to authenticate, authorize, and share his/her data. Just as non-portable telephone numbers locked customers into the carrier that issued the number, non-portable identifiers lock customers into other types of data service providers (finance, insurance, health care, education, etc.).

So a core tenet of the personal data ecosystem is that informational self-determination can only be achieved if individuals have full portability of *both* their digital identifiers and their personal data. However, this architectural requirement is much more easily stated than implemented, for two reasons:

1. **Portable data requires persistent identifiers.** If personal data is shared using identifiers that may change if an individual changes his/her legal name, domain name, telephone number, email address, or any other type of conventional address, the sharing relationship will break if the identifier changes. So none of the identifiers in wide use today meet the requirements of full lifetime identity and data portability;

26 See G. Tomko, D. Borrett, H. Kwan, G. Steffan. (August 2010). "SmartData: Make the data 'think' for itself." In *Identity in the Information Society*. Volume 3, Issue 2, pp. 343–362. Retrieved from http://link.springer.com/article/10.1007%2Fs12394-010-0047-x; I. Harvey, A. Cavoukian, G. Tomko, D. Borrett, H. Kwan, D. Hatzinakos (eds.). (2013). *SmartData: Privacy Meets Evolutionary Robotics*. Springer.

27 http://en.wikipedia.org/wiki/Local_number_portability#History

2. **Portable data requires shared semantics.** Just as banks can only interoperate by recognizing the same currencies, personal and business cloud providers can only interoperate by recognizing the same data semantics.[28] In other words, if you choose to move your personal cloud from one CSP to another, 100 percent of the data, credentials, and relationships in your personal cloud need to keep right on working. This means that data portability relies heavily on semantic data interchange—the two are inseparable.

5. Data-By-Reference (or Subscription)

Today, an organization that needs personal data to do business has no choice but to store a copy of that data if it requires repeated access—whether for ongoing use, customer service, personalization, marketing, legal and regulatory compliance, or Big Data analytics. There is simply no feasible alternative, despite the long list of drawbacks to personal data storage:

- **The cost** of building, operating, and maintaining the various databases;
- **The need to refresh the data** because it is constantly growing stale;
- **The regulatory requirements** of protecting the privacy of the stored data; and
- **The potential liability** of costly data breaches.

For these organizations, the PDE represents the first real alternative to personal data storage: *accessing the authoritative record of the data directly from the customer's personal cloud*. With the speed of modern Internet and database technologies, the "smarts" of semantic data interchange protocols (and future artificial agents sufficiently intelligent to serve as a contextual gatekeeper for one's personal data), the protection of user-centric trust frameworks, and the continuous access enabled by data portability, it will soon be entirely feasible for organizations to store a reference to the data and access it in a customer's personal cloud only when needed.

This approach is a *Privacy by Design,* win-win strategy for both individuals and organizations:

1. **Organizations save money** by reducing data storage and maintenance costs;
2. **The data is always fresh** because the organization is accessing the customer's "golden" copy;
3. **User control is dramatically increased** because the user can adjust or revoke data access rights directly at his/her personal cloud;

28 This includes the ability to map and do just-in-time translations between the data semantics used across different companies, markets, domains, and applications, since we will never have a single universal language for all data (any more than we have a single universal language for all people).

4. **Liability for data breaches is dramatically reduced** because the organization is no longer storing the data;

5. **The lifetime value of a customer grows** if the organization is granted an ongoing subscription to the customer's personal cloud; and

6. **Customers will accrue benefits**, monetary or otherwise, by granting a subscription to the organization.

As attractive as it is in the near term, data-by-reference is not always possible, either because of latency or server availability issues, or because the data is encrypted in the user's personal cloud and can only be unlocked via a private key to which only the user has access (e.g., on a local device such as a smartphone or some other safe place). In this case, a copy of the permissible portion of the personal data the user is willing to share must still be held by the subscribed business cloud (and strongly protected). However, as intelligent cloud-based gate keepers (i.e., the managers of smart data) are developed in the future, it may be imperative to minimize latency and availability issues with the protective cryptography for some use cases. This may mean the agents are trusted to hold the user's private key, or that a user's self-hosted personal cloud shares temporary keys to delegate separate temporary and limited cryptographic control with one or more agents.

However, even in the data-by-subscription scenario there are significant privacy and trust advantages:

1. **Organizations can be notified** of personal data changes in real time or near-real time.

2. **Updates can be delivered** using message-level encryption for high security.

3. **Link contracts can govern** the continued privacy and security safeguards that must be observed by the subscriber.

4. **The user can continue to exert direct control** and benefit from his/her personal cloud by the continuing subscription.

Data-by-reference and data-by-subscription are also bidirectional when organizations are the authoritative sources for some types of personal data such as government licenses, medical test results, bank account balances, and reputation or credit scores. The individual may not have direct write access to all of this data, but in many contexts has rights to control how it is accessed or used. Semantic data interchange supports the definition and management of references, rights, and subscriptions.

6. Accountable Pseudonyms

The ability of an individual to control the sharing of his/her personal data is inextricably bound with the ability of an individual to control the identifier(s) he/she shares in a particular context. Identity is as contextual as privacy; limiting a person's ability to choose a contextual identity inherently limits the privacy that person may enjoy, in that context.

Therefore another key tenet of the PDE is that, within the limits of applicable laws, organizations should offer individuals the choice to use a pseudonym instead of a verinym (an identifier associated with an individual's public identity). The privacy benefits of pseudonyms are widely extolled; one need look no further than the legal foundation of democratic nations, where voting is required to be pseudonymous in order to protect the privacy of each voter.

However, supporting pseudonymity does not mean abandoning accountability. For example, most democratic voting systems permit the pseudonymity of a voter to be "pierced" by court order to investigate voter fraud or other election irregularities. This is *Privacy by Design* at work: the voting system architecture requires that:

1. Before the vote is accepted, the voter must register an association between the pseudonymous vote and the voter's verinymous legal identity;

2. At the same time, this association must remain hidden until there is a legal order to reveal it for an investigation—and even then, it is revealed only to the investigating parties.

This same principle can be applied to identification in the PDE. For example:

- **Personal cloud service providers** can offer users the ability to register pseudonyms that are not publicly associated with the user's public identity, but which allow the association to be discovered under a court order or other process defined in the applicable trust framework;

- **User-centric trust frameworks** can also specify the de-identification requirements that must apply when accepting pseudonyms together with the re-identification requirements necessary to hold the user accountable if he/she violates the rules of the trust framework (or other legal obligations);

- **User experience designers** can make pseudonyms the default for certain contexts or permit users to select usage of a pseudonym with as little as one click—while also educating users about accountability mechanisms; and

- **Reputation systems** can be designed to aggregate reputational feedback represented by multiple pseudonyms to a single individual, while keeping this aggregation private within the reputation system until there is valid legal reason to expose it.

Like data portability, persistent accountable pseudonyms are closely tied with semantic data interchange—in fact they have specific representations in semantic graph models. This makes it possible to design pseudonyms directly into the architecture and trust frameworks for online trust networks—a perfect example of the type of systemic protection *Big Privacy* is designed to provide. By baking support for accountable pseudonymity into the fabric of the PDE, we maximize the value of pseudonymity to all who need it—including the operators of Big Data repositories for whom it may be particularly valuable.

7. Contractual Data Anonymization

While accountable pseudonyms give individuals an extra degree of privacy when sharing data, there are other contexts in which the sharing needs to be completely anonymous, i.e., unlinkable to an individual's true identity (or as close to that as technically possible).

This too is a very hard problem—equal to that of pseudonymity—but on which, thankfully, much progress has been made. Several technologies have been developed for de-identification of data sets that can make re-identification extremely difficult, if not impossible, while ensuring that the level of data quality is appropriate for the secondary purpose at issue. For example, in the health sector, Dr. Khaled El Emam, Canada Research Chair in Electronic Health Information at the University of Ottawa, has developed a framework for de-identifying personal health information in a manner that simultaneously minimizes both the risk of re-identification and the degree of distortion to the original database.[29]

In the context of Big Data, the issue with these technologies is not their effectiveness, but the control that individuals have over them. Today, that control is effectively zero—de-identification is performed entirely on the side of organizations or data brokers using the personal data. So it is little wonder that users have growing concerns about whether their personal data has been properly de-identified for usage in Big Data sets and whether re-identification may occur at any time (especially without the user's consent).

The PDE offers two new tools to enable users to wield the power of de-identification technologies:

1. **User-side de-identification.** Personal clouds truly bring "the power of the cloud" to individuals, and there is no better example than applying de-identification to data stored in or accessed through a personal cloud *before* it is shared with a third party. For example, if a user's personal cloud can access his/her Electronic Medical Records (EMRs), the user could instruct his/her personal cloud to de-identify this data according to an agreed-upon standard, prior to sharing it for a population health study;

2. **Contractual data anonymization.** Whether or not user-side de-identification is applied, if a user shares data using a semantic data interchange contract, the contract can require de-identification, conforming to certain standards or policies. *This machine-readable contract can now travel with the de-identified data*, so that all parties processing it downstream, including Big Data analytics providers, may continue to uphold the de-identification requirements.

29 See A. Cavoukian, K. El Emam. (2010). "A Positive-Sum Paradigm in Action in the Health Sector." Retrieved from http://www.ipc.on.ca/images/Resources/positive-sum-khalid.pdf; K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, J. Bottomley. (2009). "A Globally Optimal k-Anonymity Method for the De-identification of Health Data," in *Journal of the American Medical Informatics Association*, 16(5): 670–682. Retrieved from http://jamia.bmj.com/content/16/5/670.

In terms of the "tools and rules" of user-centric trust frameworks, contractual data anonymization is a privacy-respecting *tool* that can be matched with a policy *rule* requiring trust framework participants to conform to the user's selected de-identification policy throughout the life cycle of the data (whether for a Big Data life cycle or any other usage). This is a textbook example of how users can exert the radical personal control that is the very hallmark of *Big Privacy*.

5. Exemplar: Respect Network™ and the OASIS XDI Protocol

The seven architectural elements of *Big Privacy* are currently being implemented by many different contributors to the PDE. In this section we will focus on one specific example: how Respect Network is using the OASIS XDI protocol to build a global private data sharing network that incorporates all seven elements.

1. The Respect Network

The Respect Network is an alliance of 33 companies and organizations[30] working together to build the world's first personal cloud network.[31] Representative of all segments of the PDE, the Founding Partners of the Respect Network currently fall into eight categories:

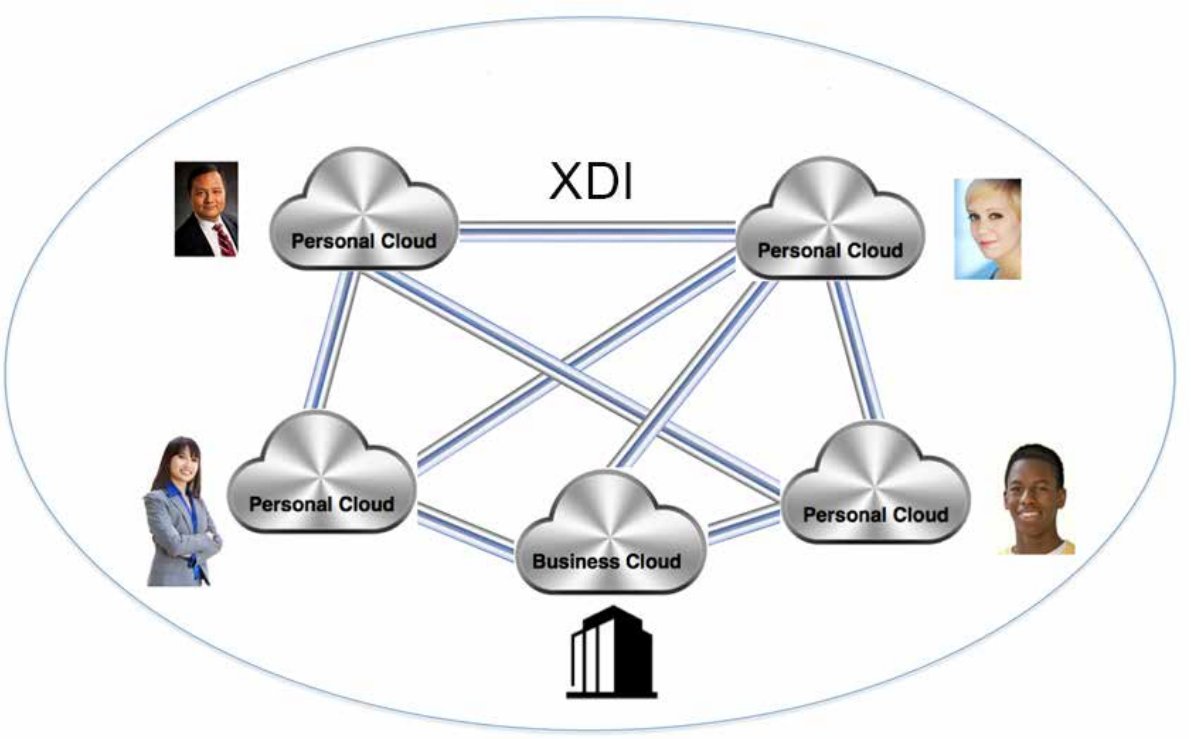
1. Commercial Partners
2. Ecosystem Partners
3. Cloud Service Provider (CSP) Partners
4. Development Partners
5. Verification Partners
6. Integration Partners
7. Non-profit Partners
8. Consulting Partners

At the European Identity Conference in May 2011, the Founding Partners announced their collaboration to create a globally-scalable, peer-to-peer (P2P) network of personal and business clouds. Any two clouds on the Respect Network may form connections and securely share data and messages using the open standard OASIS XDI semantic data interchange protocol (see the next element).[32]

30 http://respectnetwork.com/founding-partners/ New Founding Partners continue to join each month.

31 http://respectnetwork.com/what-is-respect-network/

32 http://www.oasis-open.org/committees/xdi/



In many ways the Respect Network resembles the global email network, where: a) all inboxes have an email address, b) every network endpoint is discoverable via the Domain Name System (DNS), and c) email messages may be exchanged directly between any two peers. On the Respect Network: a) all clouds have an XDI address, b) every network endpoint is discoverable via the XDI discovery protocol, and c) XDI messages can be exchanged directly between any two peers. In both cases there is no centralized middleman that has access to all the messages and all the data—in effect, every personal and business cloud is practicing "informational self-determination."

2. The OASIS XDI Semantic Data Interchange Protocol

When the XDI Technical Committee at OASIS (Organization for the Advancement of Structured Information Standards) was formed in 2004, its mission was to define a standard format and protocol for sharing *semantic data*. This means much more than putting data into a structured markup language like XML (Extensible Markup Language)[33] or structured format like JSON (JavaScript Object Notation).[34] It means describing the data using a standard semantic graph model so that machines can "read" the data and make inferences about it.

A simple example is two common forms of identification: business cards and driver's licenses. Data schemas for business cards, such as vCard, have existed

33 https://en.wikipedia.org/wiki/XML

34 https://en.wikipedia.org/wiki/JSON

since at least 1995,[35] and for driver's licenses (in the United States) since 2005.[36] However, if a machine is given one vCard and one electronic driver's license and then asked if they describe the same person—a trivial task for a human—the machine would find it extremely difficult to do. This is because vCards and electronic driver's licenses use different data schemas.

With semantic data formats, the same semantic statements describing the data ("this is a person's first name," "this is a person's last name," "this is a person's date of birth") may be shared across many different data schemas (business cards, driver's licenses, insurance records, electronic medical records)—*even if they are in different human languages*. So if a machine is asked whether a semantic electronic business card and a semantic electronic driver's license describe the same person, it can answer almost as accurately as a human.

The goal of the XDI Technical Committee is not just a semantic data *format*, but a semantic data *protocol* that enables machines to literally "talk" to each other in a common language, i.e., to be able to ask and answer questions about data, just as people do. For example:

Machine 1: "Do you know the age of this person?"

Machine 2: "Yes, but I cannot give you the exact date."

Machine 1: "Is she old enough to drive in Michigan?"

Machine 2: "Yes."

From a privacy perspective, the most powerful feature of XDI as a semantic data interchange protocol is that it can use semantic statements to describe *the rights and permissions that apply to a specific set of data in a specific context*. The result is a revolutionary new approach to data control called a **link contract**. An XDI link contract functions just like a real-world contract except that is readable by a machine and can include policy expression statements that are enforceable by machines. So once an individual approves a link contract granting a set of permissions to his/her personal data (e.g., to read it, modify it, copy it, move it, reshare it, delete it), that individual's personal cloud can then automatically enforce those permissions.

Most importantly, since the link contract itself is written in XDI, *these permissions can travel with the data*. This means that any other system that requires a copy of the data can also have a machine-readable copy of the link contract that covers it. So we can finally start using the growing intelligence of our machines and networks to enhance our privacy instead of leading to its erosion.

In the future, this methodology can extend even further. Using a technology referred to as "SmartData,"[37] an intelligent agent will empower personal data by wrapping it in a "cloak of intelligence" such that it now becomes an individual's

35 https://en.wikipedia.org/wiki/VCard

36 https://en.wikipedia.org/wiki/REAL_ID_Act

37 See G. Tomko, D. Borrett, H. Kwan, G. Steffan. (August 2010). "SmartData: Make the data 'think' for itself." In *Identity in the Information Society*. Volume 3, Issue 2, pp. 343–362. Retrieved from http://link.springer.com/article/10.1007%2Fs12394-010-0047-x; I. Harvey, A. Cavoukian, G. Tomko, D. Borrett, H. Kwan, D. Hatzinakos (eds.). (2013). *SmartData: Privacy Meets Evolutionary Robotics*. Springer.

virtual proxy in cyberspace, controlling the release of his/her data in accordance with the user's preferences. As a result, personal data can be stored and shared as a constituent of the binary string specifying the intelligent agent. Such an agent would proactively build in privacy and security right from the outset, so that nothing is treated as an afterthought. It would embody a foundation of control and trust within the technology itself as the first line of defense, incorporating the principles of purpose specification, personal consent, security, and use limitation.

3. The Respect Trust Framework

The inspiration for the Respect Network was the idea of creating a trust framework "of the people, by the people, for the people," i.e., where the tools and rules are designed from the outset to uphold the 7th principle of *Privacy by Design*: **Respect for Users—Keep It User-Centric**.

To do this, Respect Network legal architect Scott David worked with the Respect Network founders to condense Fair Information Practice Principles (FIPPs) from around the world into five core principles, called the **Respect Principles**. These are the rules to which all members of the network (both individuals and businesses) must agree. The five principles are presented in their entirety below:

Principle	Synopsis	Wording
1. Promise	*We will respect each other's digital boundaries*	Every Member promises to respect the right of every other Member to control the identity and personal data they share within the network and the communications they receive within the network.
2. Permission	*We will negotiate with each other in good faith*	As part of this promise, every Member agrees that all sharing of identity and personal data and sending of communications will be by permission, and to be honest and direct about the purpose(s) for which permission is sought.
3. Protection	*We will protect the identity and data entrusted to us*	As part of this promise, every Member agrees to provide reasonable protection for the privacy and security of identity and personal data shared with that Member.
4. Portability	*We will support other Members' freedom of movement*	As part of this promise, every Member agrees to ensure the portability of the identity and personal data shared with that Member.
5. Proof	*We will reasonably cooperate for the good of all Members*	As part of this promise, every Member agrees to share the reputation metadata necessary for the health of the network, including feedback about compliance with this trust framework, and to not engage in any practices intended to game or subvert the reputation system.

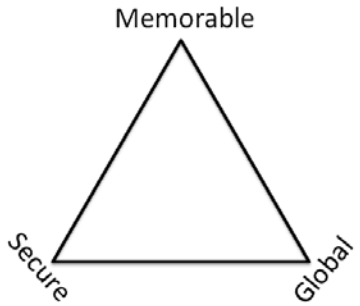
The final principle, **Proof**, is the basis for the other core innovation of the Respect Trust Framework: the entire Respect Network functions as a P2P reputation system, in which every member who has a verified relationship with another member may provide peer reputational feedback. This **Respect Reputation System** creates a strong incentive for all members to uphold the Respect Principles. It was first implemented by the Connect.Me socially-verified reputation service which won the highly esteemed Privacy Award at the 2011 European Identity Conference.[38]

The Respect Trust Framework was listed as an open public document with the international non-profit Open Identity Exchange (OIX) in May 2011.[39] It is not intended to be a "one-size-fits-all" policy tool. Rather, it is designed to be a top-level "umbrella" trust framework that establishes the smallest and most universal set of policies that can apply to all members of the trust network. Other more specialized trust frameworks can then "plug in" to specify more detailed technical and legal requirements that may apply in specific legal jurisdictions or other trust domains (e.g., health care, finance, education, etc.).

4. Cloud Names and Cloud Numbers

The fourth architectural element of *Big Privacy*—identity and data portability—is one of the most difficult to implement because of a well-known problem with the very nature of identifiers and identification known as Zooko's Triangle:[40]

Zooko's Triangle essentially says that when it comes to identifiers on a digital network: "Memorable, Secure, Global—pick any two." In other words:



- **Identifiers that are memorable and secure** (like human names) are not globally unique;
- **Identifiers that are memorable and globally unique** (like domain names) are very hard to secure (because it is easy to trick people into using the wrong ones);
- **Identifiers that are secure and globally unique** (like very long numbers) are impossible for people to use.

38 http://blog.connect.me/connectme-wins-2011-eic-privacy-award

39 http://openidentityexchange.org/trust-frameworks/respect-trust-framework/

40 https://en.wikipedia.org/wiki/Zooko%27s_triangle

Zooko's Triangle is an especially hard problem for the PDE for the following reasons:

a) To protect people's personal data, identifiers MUST be secure;

b) To achieve lifetime data portability, identifiers MUST be globally unique AND never need to change (persistent);

c) To be usable by people everywhere, identifiers MUST be memorable.

In order to solve Zooko's Triangle, the OASIS XDI Technical Committee leveraged the power of semantic data description. First, they created two types of identifiers to be used natively in XDI:

1. **Cloud names** are simple, human-friendly identifiers like =forrest.gump intended to be very easy for people to remember and use—much like domain names, only even simpler;

2. **Cloud numbers** are long strings of letters and numbers known as UUIDs.[41] They are globally unique and never need to change, but are also impossible for ordinary mortals to use—for example, [=]!:uuid:f81d4fae-7dec-11d0-a765-00a0c91e0001.

Second, they created a way for XDI to reuse almost any other popular identifier (email address, phone number, URL, IP address, etc.) using a syntax called a **cross-reference**. For example, here is what an email address looks like when used as an XDI cross-reference to a person:

=(mailto:forrest.gump@example.com)

Third, they created a standard semantic statement for how *one identifier can point to another*. This is how cloud names and cross-references (which are human-memorable and globally unique) can point to cloud numbers (which are secure and globally unique—and persistent). For example:

=forrest.gump/$ref/[=]!:uuid:f81d4fae-7dec-11d0-a765-00a0c91e0001

=(mailto:forrest.gump@example.com)/$ref/[=]!:uuid:f81d4fae-7dec-11d0-a765-00a0c91e0001

By bringing together three types of identifiers—cloud names, cloud numbers, and cross-references—the XDI Technical Committee "squared Zooko's Triangle." This means that XDI—as an open standard supported by all Respect Network members—can satisfy all three requirements of identification in the PDE—human memorability, security, and global uniqueness and portability. This is why the Respect Network can standardize on these three types of identifiers for all Respect Network personal and business clouds.

41 http://en.wikipedia.org/wiki/UUID

5. Respect Connect and Personal Channels

Social login—the ability for users to register and/or log in at a website using their social networking account—has become a major feature of social networks like Facebook, Twitter, LinkedIn, and Google+. It has become so popular with websites that, as of August 2013, Facebook alone was providing over 850 million social logins per month.[42]

However, this practice has been raising growing privacy concerns. To begin with, users are sharing all their social login activity with a social network whose terms of service gives it very broad rights to use this information for any purpose. Secondly, the user's social login information is not portable—it is locked into the social network he/she uses for each login. Thirdly, both users and sites using social logins are building in a growing dependence on the social network as the middleman in every relationship.

A personal cloud network like Respect Network can solve these problems. The first major feature the Founding Partners are bringing to market is Respect Connect "safe single sign-on."[43] It works almost identically to a social login—users just click the Respect Connect button at a website to register and/or login—but the result is *a direct P2P connection with the user's own personal cloud*. From a privacy standpoint, the advantages are overwhelming:[44]

1. **The user's login information remains completely private** and under his/her personal control—there is no middleman with access to all the shared data.

2. **The user's login information is fully portable for life**, protected by the portability principle of the Respect Trust Framework, no matter how often the user changes his/her cloud service provider or self-hosted personal cloud server.

3. **All data sharing is under the XDI open standard**, so the user and the site are not locked in to any proprietary protocols or data schemas.

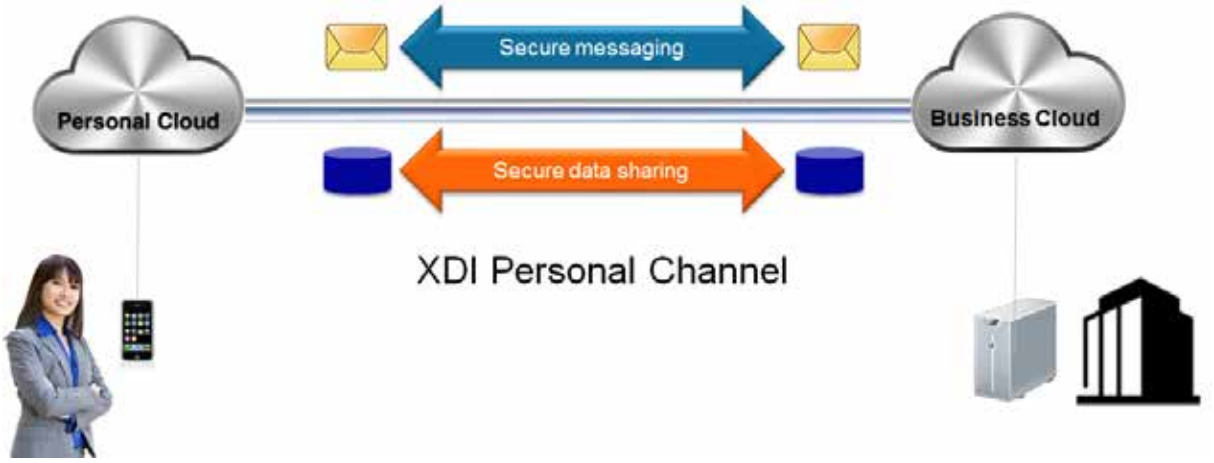
4. **Users and sites can agree on precisely the information a particular individual is willing to share in a particular context**, and do not need to be subject to the limitations or constrains imposed by a middleman's policies or protocols.

But the real power of Respect Connect goes far beyond social login. Every Respect Connect button is an invitation to create a secure, private, persistent peer-to-peer XDI connection between two personal clouds—or between a personal cloud and a business cloud. This new form of intelligent communications link is called a *personal channel*.

42 http://mashable.com/2013/10/29/facebook-dominates-social-logins/

43 A term coined by VRM (Vendor Relationship Management) pioneer Doc Searls at the May 2011 Internet Identity Workshop in Mountain View, CA.

44 For more details, see http://respectnetwork.com/webinars/respect-connect/



What is unique about XDI personal channels is that they can do both *secure messaging*—the semantic equivalent of email, SMS, or instant messaging—and *secure data sharing*—the semantic equivalent of the file, photo, and status sharing popular on social networks and file sharing services like Dropbox, Google Drive, and Microsoft Skydrive. But in both cases the security, privacy, portability, and persistence (if desired) is much stronger because on the Respect Network, all cloud-to-cloud connections use XDI link contracts anchored in the strong personal privacy protections (both legal and technical) contained in the Respect Trust Framework.[45]

It is these protections that finally give websites the choice to no longer store personal data if it is not absolutely necessary to do so; rather, they can use data-by-reference (to access the current authoritative record of that data in a user's personal cloud over a permissioned personal channel) or data-by-subscription (to receive secure, authenticated updates when that data changes—for example, when an individual is moving, changing jobs, changing clothing sizes, getting a mortgage, etc.).

6. Respect Pseudonyms

Systemic control of personal data starts with systemic control of personal identifiers. So if the PDE is going to support persistent, accountable pseudonyms, this capability must be proactively designed in, at the system architecture level. This is why it has been a design goal from the inception of both the Respect Network and the OASIS XDI Technical Committee.

As explained above, the primary challenges in supporting accountable pseudonyms are:

1. Making them easy for both individuals and sites to use;

45 For more details see the Respect Network white paper, *The Personal Channel: The Extraordinary Benefits of Communicating Via Personal Clouds*, http://respectnetwork.com/papers/

2. Enabling them to be portable and persistent;

3. Creating incentives for all parties in the PDE to respect the user's right for the pseudonym not to be correlated with the user's public identity unless there is a valid legal reason to do so.

To meet these challenges, the Respect Network uses a combination of the technical capabilities of XDI and the policy capabilities of the Respect Trust Framework.

For example, to make pseudonyms easy to use, support for their use is built directly into XDI semantics. Rather than the **$ref** ("reference") relation discussed previously, which is a publicly discoverable correlation between one identifier and another (e.g., a cloud name to a cloud number), the **$rep** ("replacement") relation allows one identifier to *privately* point to another, i.e., create a correlation that is known to the user but not to anyone else interacting with the user's XDI graph. (In fact, to the outside world, it will look like an entirely separate personal cloud.)

This means, for example, that when using Respect Connect to create a new personal channel, the user can easily choose an existing pseudonym or instruct his/her primary personal cloud to register and share a new pseudonym.

Since all pseudonyms on the Respect Network are cloud names and cloud numbers as described above, they retain all the usability, privacy, portability and persistence benefits of cloud names and cloud numbers. Persistence is particularly important for pseudonyms intended for long-lived, privacy-protected engagement in a community, such as is often needed for academic or political discourse and dissent.

Finally, the Promise Principle of the Respect Trust Framework —"*to respect the right of every other Member to control the identity and personal data they share within the network*"—is the highest level policy establishing the obligation of all Respect Network members (individuals or businesses) to respect the right of other members to use accountable pseudonyms where legal identity is not required. This extends to the design of the Respect Reputation System, where it is critical that a Respect Pseudonym be held accountable for both positive actions that add to its owner's reputation and negative actions that diminish it.

7. Anonymous Link Contracts

To implement contractual data anonymization, the Respect Network employs XDI link contracts together with a core feature of the XDI semantic graph model called **contextualization**. This is the ability to take any XDI graph of data—such as a set of personal health measurements—and place it into a different XDI context that gives it new meaning (and subjects it to new operational policies).

In the case of data anonymization, the context is **$anon**, the special context defined by the XDI Technical Committee for data to be shared under a link contract that requires de-identification. So, for example, when a person's demographic profile is shared in the $anon context, the receiving party is contractually agreeing to: a) de-identify the data (if that has not already been done by the user's personal cloud—see **user-side de-identification** above), and b) keep the data de-identified

unless the user provides explicit opt-in consent (such as to be notified about a possible serious health risk).

Not only does a link contract with a $anon provision codify a legal requirement for the receiving party to not re-identify the data, but—as we have stressed throughout this paper—with XDI this link contract can travel along with the data. So if the data is re-shared (with permission), this obligation can then be passed on to be enforced by other data analytics providers downstream.

This is the kind of systemic data protection that is required, for example, to make the patients in a large health-care network comfortable with having their personal health-care records included in a large-scale Big Data study of population health. There is no question of the benefits that such analytics can bring—including insights that may be specifically beneficial to participants in the study—but these benefits can only be realized if patient's privacy concerns are adequately addressed. This is the doubly-enabling Big Win of Big Data + *Big Privacy*!

6. How *Big Privacy* Applies the 7 Foundational Principles of *Privacy by Design*

Big Privacy is a further development and extension of the concept of *Privacy by Design*. *Privacy by Design* (*PbD*) is an internationally recognized[46] framework that addresses the ever-growing and systemic effects of information and communications technologies and large-scale networked data systems. *PbD* advances the view that the future of privacy cannot be assured solely by compliance with regulatory frameworks; rather, privacy assurance must ideally become an organization's default mode of operation. The objectives of *PbD*—ensuring privacy and gaining personal control over one's information and, for organizations, gaining a sustainable competitive advantage—may be accomplished by practicing the following 7 Foundational Principles:

1. *Proactive* not Reactive; *Preventative* not Remedial
2. Privacy as the *Default Setting*
3. Privacy *Embedded* into Design
4. Full Functionality – *Positive-Sum*, not Zero-Sum
5. End-to-End Security – *Full Lifecycle Protection*
6. *Visibility* and *Transparency* – Keep it Open
7. *Respect* for User Privacy – Keep it *User-Centric*[47]

In response to the unique privacy challenges posed by Big Data, *Big Privacy* applies the 7 Foundational Principles of *Privacy by Design* to provide systemic protection

46 See International Conference of Data Protection and Privacy Commissioners, Jerusalem, Israel. (October 27–29, 2010), "Resolution on *Privacy by Design*." Retrieved from http://www.justice.gov.il/NR/rdonlyres/F8A79347-170C-4EEF-A0AD-155554558A5F/26502/ResolutiononPrivacybyDesign.pdf; U.S. Federal Trade Commission. (2012). *Protecting Consumer Privacy in an Era of Rapid Change*. Retrieved from http://www.ftc.gov/os/2012/03/120326privacyreport.pdf; European Commission. (2012). "General Data Protection Regulation." Retrieved from http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

47 See A. Cavoukian. (2011). "Privacy by Design. The 7 Foundational Principles." Retrieved from http://www.ipc.on.ca/images/Resources/7foundationalprinciples.pdf

of personal data and sustained individual control over how it is collected, used, and disclosed in the PDE. In this section we will discuss how the 7 architectural elements of *Big Privacy* apply the 7 Foundational Principles of *Privacy by Design* to the unique privacy challenges of Big Data in the PDE.

1. Proactive not Reactive; Preventative not Remedial

The *Privacy by Design* approach is characterized by proactive rather than reactive measures. It anticipates and prevents privacy-invasive events before they happen. *PbD* does not wait for privacy risks to materialize, nor does it offer remedies for resolving privacy infractions once they have occurred—it aims to **prevent** them from occurring. In short, *Privacy by Design* comes before the fact, not after.

The trend in networked systems and information and communications technologies has driven towards a centralized model of personal information storage where one company sets all the terms and has access to all the data—in effect, creating a single giant information silo where individuals are effectively cut off from meaningfully participating in an economy based on a resource that they themselves produce, namely their personal information. Privacy risks abound under such a model.

Personal clouds invert the paradigm of information silos by creating a new kind of network in which all actors are "peers" interacting on a level playing field where all connections—be they person-to-person, person-to-business, or business-to-business—are peer-to-peer, with no "middleman" in between. In other words, in contrast to the predominant social-media model, *no one has access to all the data.* Individuals are empowered to own and control their own data, and to use and share it with other participants in the PDE on terms that *they* set and negotiate, as need be.

Personal cloud networks with trust frameworks and semantic data will enable individuals not only to control the initial release of their data, but to also control its future life cycle use within Big Data and other systems. While many systems put a premium on user consent at the time of the access request, personal cloud networks make it possible for users to configure access policies before a requesting party ever interacts with their personal data or resources. This enables a different relationship between users and applications (or even other autonomous parties) seeking access: the owner now offers terms instead of being asked to accept them. This helps to correct the "power imbalance" between individuals and online service operators, and between client software and server software.

2. Privacy as the Default Setting

We can all be certain of one thing—the default rules! *Privacy by Design* seeks to deliver the maximum degree of privacy by ensuring that personal data is automatically protected in any given IT system or business practice. If an individual does nothing, their privacy still remains intact. No action is required on the part of the individual to protect their privacy—it is built into the system, by default.

Under the model of centralized storage of personal information, privacy is defined by privacy policies. While these are meant to promote openness and greater transparency of an organization's processing of personal information, the vast majority of privacy policies, in fact, do precisely the opposite—they are typically long, unclear, difficult to understand, and present terms and conditions on a "take it or leave it" basis. As a result, organizations are increasingly using them, not as a means of promoting openness and transparency in their processing of personal information, but rather as a tool of more aggressive personal information collection practices.

In stark contrast, a trust framework is a network-wide document that legally binds *all* members of a trust community—both individuals and organizations—to a set of business, legal, or operational policies, as a condition of membership. A user-centric trust framework sets down global terms and conditions for interacting with personal clouds in a manner that respects the privacy of individuals, with strong assurances of security. In this way, individuals no longer need to worry about whether their privacy is being respected in their interactions with organizations or other individuals. Instead, when operating under a trust framework compatible with *Big Privacy*, individuals can be confident that their privacy is being protected because it is automatically built into the system, by default.

3. Privacy Embedded into Design

Privacy by Design is embedded into the design and architecture of IT systems and business practices. It is not bolted on as an add-on, after the fact. The result is that privacy becomes an essential component of the core functionality being delivered. Privacy is integral to the system, without diminishing functionality.

The very definition of *Big Privacy* is embedding privacy into the design of the PDE and how it can be used in the context of Big Data. The PDE is explicitly architected as a network of peer-to-peer connectivity over private personal channels that avoid both information silos and unnecessary "middlemen" between interactions. Connections are mediated by machine-readable link contracts that enforce the preferences and permissions a user has provided with regard to shared data. These link contracts are also designed to travel alongside the data sent to a receiving system so that the user's preferences and permissions can continue to be enforced for the complete life cycle of the data. In addition, a personal cloud network can also provide data-by-reference, accountable pseudonymity, and contractual data anonymity, bringing ground-breaking personal privacy features to thousands or millions of relying parties so they can meaningfully share and use data fully under the user's control.

4. Full Functionality – Positive-Sum, not Zero-Sum

Privacy by Design seeks to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner, not through a dated, zero-sum approach, where unnecessary trade-offs are made. *Privacy by Design* avoids the pretense of false dichotomies, such as privacy vs. security, demonstrating that it is indeed possible to have both.

A recent Edelman industry study[48] found that consumers have never been more concerned about the security of their personal information and believe that businesses are mismanaging privacy and security issues. As a result, the study claims that businesses and organizations must do more to better educate themselves at handling privacy and security issues—or else they will lose customers' trust—and with it, their business.

This study supports a claim which proponents of *Privacy by Design* have been making for many years, namely that privacy should be viewed, not as a compliance issue, but rather as a *business issue*. Privacy, in other words, is good for business. It facilitates continuation of valuable business relationships, serves to preserve existing customers and attract new ones, and fosters the development of a sustainable competitive advantage. In short, privacy builds consumer confidence and trust.[49]

In this way, the PDE should not be viewed as a burden to the private sector, but rather as a boon to it! When users are fully aware—because they are fully in control—of the privacy implications of their data-sharing decisions, they will not only have more confidence and trust in making those decisions, but will be more willing to share their personal information with trusted companies. This will, in turn, stimulate the online economy and benefit those companies that treat individuals' personal information in a respectful manner.

Big Privacy grants users greater control over their personal information by offering greater choices in terms of providers for identifiers, identity credentials, and personal clouds. Through identity and data portability, personal identities and data are made portable across sites and service providers in the PDE. Trust frameworks can bind every member of a personal cloud network to follow fair information privacy practices (including providing portability of core data, personal data, policies and relationships), for mutual benefit and protection. In this way, *Big Privacy* helps to accommodate all legitimate interests and objectives in a positive-sum "win-win" manner by providing a trustworthy platform of interoperability built on maximizing user confidence and trust, thereby maximizing users' willingness to share their personal information. As such, both individuals and creators/owners of Big Data algorithms will benefit.

5. End-to-End Security – Full Lifecycle Protection

Privacy by Design, having been embedded into the system prior to the first element of information being collected, extends security throughout the entire life cycle of the data involved. This ensures that all data is securely retained, and then securely destroyed at the end of the process, in a timely fashion. Thus, *Privacy by Design* ensures cradle-to-grave, secure life cycle management of information, end-to-end.

48 See Edelman. (2012). Privacy & Security. The New Drivers of Brand, Reputation and Action. Retrieved from http://datasecurity.edelman.com/wp-content/uploads/2012/03/Data-Security-Privacy-Executive-Summary.pdf.

49 See A. Cavoukian, T. Hamilton. (2002). *The Privacy Payoff: How Successful Businesses Build Consumer Trust*. Toronto: McGraw-Hill Ryerson.

Strong security measures undertaken by a single person, service provider or business online, become meaningless if all parties in the interaction do not have compatible policies, practices and technologies. Weak links in security can lead to data breaches, compromised identity credentials, and identity theft. There is a need to proactively ensure privacy within any federated identity system, given the privacy issues arising from data-in-motion between multiple stakeholders.

Personal cloud networks apply protection at the interface of the user and requesting party; other back-end systems surrounding the resources in question are responsible for protecting them during their entire life cycle.

Personal information that has reached the end of its life cycle must be destroyed in a consistently secure and privacy-protective manner. Not replicating personal information is a powerful way of keeping it more secure throughout its life cycle, since it prohibits the existence of multiple copies. In the PDE, personal clouds are the authoritative source for personal information, and businesses may gain access to it through a "subscription" model. Through the use of data-by-reference, access to personal information can be terminated when the individual chooses to delete the data or terminate the relationship. In this way, *Big Privacy* ensures that all personal information is securely transmitted, retained, and then destroyed at the end of its life cycle.

6. Visibility and Transparency – Keep it Open

Privacy by Design seeks to assure all stakeholders that whatever the business practice or technology involved, it is in fact operating according to the stated promises and objectives, subject to independent verification. Its component parts and operations remain visible and transparent, to users and providers alike. Remember—trust but verify!

A core premise of the PDE is that users must be given a clear understanding of the overall context and usage of their personal data, including access and sharing by any party (including the user's own Cloud Service Provider). This is currently the only way to achieve a real sense of control over personal data that will grow over time.

User-centric trust frameworks are the primary *Big Privacy* mechanism that can provide this kind of system-wide transparency. Because they are published for public review and scrutiny, and because they are binding on all members of a trust network that agree to operate under them, they can establish privacy-respecting practices that become the norm across broad communities of usage.

Trust framework agreements must also define the assessment, accreditation, and enforcement mechanisms through which individuals and other stakeholders can verify and ensure that all members are following fair information privacy practices. For example, a contextual reputation system, provided as part of a personal cloud network, can harness the "wisdom of crowds" to observe members on an ongoing basis, thereby providing them with a powerful incentive to follow the rules established by the trust framework.

7. Respect for User Privacy – Keep it User-Centric

Above all, *Privacy by Design* requires architects and operators to keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, and empowering user-friendly options. Keep it user-centric!

At its core, respecting the user means that, when designing or deploying an information system, the individual's privacy rights and interests are accommodated right from the outset. User-centricity is anticipating and designing in a person's privacy perceptions, needs, requirements, and default settings.[50] It means putting the interests, needs, and expectations of people first, not those of the organization or its staff. Empowering people to play active roles in the management of their personal data helps to mitigate abuses and misuses. Adding a contextual reputation system enabling public feedback when abuses or misuses occur may be an even more powerful check and balance.

The evolution of the PDE is by its very nature aligned with the 7th Foundational Principle of *PbD*, precisely because of its focus on user control. The principle of "Respect for User Privacy – Keep it User-Centric" requires that architects and operators keep the interests of the individual uppermost by offering such measures as strong privacy defaults, appropriate notice, empowering user-friendly options and enabling mechanisms for user feedback. *Big Privacy* takes user-centricity to the next level by offering users the convenience and control of a personal cloud; the protection of a user-centric trust framework; the freedom of identity and data portability, the advantage of maintaining a pseudonym, rather than only a public identity; and the ability to share personal data under a link contract that binds the relying party to maintain anonymity of that data. All of these advancements mean that *Big Privacy* will produce a paradigm shift in privacy from an "organization-centric" to a balanced model which is far more user-centric.

7. Conclusion

Privacy challenges abound in the modern world, especially in Big Data environments that collect masses of personal information and subject them to machine learning, natural language processing, signal processing, simulation, time series analysis, visualization and other analysis. Some algorithms are actually inferred in this fashion from the data itself, making it difficult to keep any privacy or use limitation guarantees to the individual, or to anticipate how PII would be used, and then obtain informed consent before collection. Big Data systems themselves generally have no solution to this dilemma and treat privacy as an externality. However, a Personal Data Ecosystem (PDE) of companies and organizations that believe individuals should be in control of their personal information is emerging, and the PDE is making available a growing number of tools and technologies to enable this control.

The *Privacy by Design* framework must be applied jointly to the PDE and to Big Data in order to achieve positive-sum, "win-win" solutions in which individuals

50 P. McDougall. (August 8, 2012). "Microsoft IE 10 Makes 'Do Not Track' Default." *InformationWeek*.

maintain control over their personal information flows, without diminishing system functionality. *Privacy by Design* can be applied at a large scale to individuals' online interactions by bringing together the user-centric architecture of the PDE with the analytic power of Big Data, resulting in a much larger "win-win" scenario, where privacy and Big Data may coexist, in tandem. This new result is what we call *Big Privacy*.

Big Privacy pushes past promise-making towards proactive data protection and elevates compliance from a piece of paper to a network-wide trust framework, with real time notification, enforcement and feedback mechanisms. *Big Privacy* transforms the debate about Big Data by making privacy a user-centric setting where the individual is empowered to choose not only what information he/she wishes to share, but also the ways in which it may be shared and under what terms. In this way, individuals may choose to allow creators/owners of Big Data algorithms to harvest and mine their data *only if* it remains de-identified, through the application of appropriate re-identification risk measurement procedures.

Big Privacy is *Privacy by Design* writ large, i.e., it is the application of the seven principles of *Privacy by Design* not just to individual organizations, applications, or contexts, but to **entire networks, value chains, and ecosystems,** especially those intended to produce and use Big Data. The goal of Big Privacy is **systemic protection** of personal data and **radical personal control** over how it is collected and used. This means that it must be possible to assure whole populations that their privacy is being respected because the network, value chain, and/or ecosystem producing and processing Big Data has implemented *Privacy by Design,* at a system-wide level.

Big Privacy comprises the following seven architectural elements which mutually complement and reinforce one another, while also enabling Big Data to coexist with privacy objectives:

- **Personal clouds** linked into **personal cloud networks** manifest a real, active PDE; they provide individuals with control of virtual compute capabilities which proactively protect personal information and engage as peers with other personal clouds or business clouds on the individual's terms;
- **Semantic data interchange** gives individuals fine-grained information sharing control and enables personal cloud services to attach individuals' privacy preferences and policies to their data in a standard, interoperable, and machine-readable form;
- **Trust frameworks** provide transparent, open governance of personal cloud network ecosystems where individuals, organizations, and service providers are members, contractually binding them to respect the rules and tools established by the trust framework;
- **Identity and data portability** provides the ultimate guarantee that individuals and organizations—not their service providers—control their own data;

- **Data-by-reference** (or subscription) enables individuals or organizations to change their minds about how their data may be used—for example, by revoking the rights of a Big Data system to analyze their data;
- **Persistent accountable pseudonyms** allow individuals to express themselves freely but with a certain discretion, remaining within the context of what is legally acceptable;
- **Contractual data anonymization** provides a way, along with accountable pseudonyms, for valuable Big Data systems to operate in compliance with all privacy regulations and personal preferences, allowing patterns to be found on an aggregate level without the need to for identifiable personal data.

In sum, this paper has reviewed the concepts of Big Data, the PDE and *Big Privacy*. It has proposed an architectural approach based on *Privacy by Design* for their coexistence. It has also presented an exemplar of the solution in the form of the Respect Network, which is presently building a *Big Privacy* architecture based on the OASIS XDI semantic data interchange protocol. As this and other solutions arrive, they will enable a positive-sum, win-win strategy in which business, governance and Big Data may all advance in the context of human dignity, privacy, and control—the Big Win of *Big Privacy*.

Office of the Information and Privacy Commissioner, Ontario, Canada
2 Bloor Street East
Suite 1400
Toronto, Ontario
Canada M4W 1A8
Telephone: 416-326-3333
Fax: 416-325-9195
E-mail: info@ipc.on.ca
Website: www.ipc.on.ca

Respect Network Corporation
3145 Geary Blvd.
Suite 419
San Francisco, CA
94118 USA
Telephone: 801-550-6695
E-mail: info@respectnetwork.com
Website: www.respectnetwork.com

Web site: www.ipc.on.ca
Privacy by Design: www.privacybydesign.ca

December 2013

Information and Privacy Commissioner, Ontario, Canada

Respect Network