The Future of Social is Personal: The Role of Personal Data Stores in Social Interfaces

Max Van Kleek and Kieron O'Hara

1 Introduction

A key characteristic common to the various kinds of "social intelligence" described in this volume is one of enhanced autonomy through technological support. Such autonomy allows constituents of a society to form new connections with others dynamically as needed, promoting a more adaptive, flexible and robust social fabric than those of traditional structures, in which efficiency leads a majority to rely on a handful of central, fixed intermediaries. This observation immediately prompts the question of whose interests that "efficiency" is designed to benefit — the intermediaries' or the users'.

While we see technology being applied in many contexts to generalise the benefits and enhance the autonomy thus described, the storage of personal information is one area where it has, thus far, been used to power a perverse reversal towards more centralisation. Currently, a handful of dominant platform vendors and application service providers are grappling for control over individuals' personal information, trying to accumulate as many users as possible in order to maximise understanding of every nook and corner of social interaction — a relentless process satirised in Dave Eggers' 2013 novel *The Circle*, about a company with the totalising slogan "All That Happens Must Be Known". This centralising trend, backed by a surveillance-and-analytics business model, began with the rise of so-called "Web 2.0", in which sites became sophisticated apps and content-management platforms designed to facilitate the creation and sharing of user-generated data and content. That content began as a few social network profiles and blog posts, but gradually

Max Van Kleek
University of Southampton, University Road, Southampton, SO17 1BJ, e-mail: emax@ecs.soton.ac.uk

Kieron O'Hara
University of Southampton, University Road, Southampton, SO17 1BJ, e-mail: kmo@ecs.soton.ac.uk

grew to encompass the entirety of personal data people keep *or generate*, from files and documents to film and music archives. Thus began a migration of personal digital artefacts from individually-administered personal computers into various information spaces of the web. The assimilation of personal data from personal digital devices has accelerated as Web application and service providers have started to create deep integrations with personal computing devices such as Facebook Home[1], Windows Skydrive[2] and Apple's iCloud[3]. Such services have extended the reach of Web services into the intimate digital spaces of one's personal devices, offering backup and management services for these private data collections as well.

What are the implications of this centralisation? Although the ultimate, long-term implications of this shift are not yet fully understood, several immediate consequences are apparent. Fundamentally, the delegation of responsibility for management of one's personal information to third party service providers necessitates relinquishing control over various aspects of how these data are handled and processed, ranging from how they are stored and represented, to how (and when) they can be accessed, as well as to whom access is granted. When third party delegation accidentally-on-purpose serves the increasingly pervasive business model of deriving revenue directly from these data themselves (through targeted advertising or licensing to third parties), platforms are essentially incentivised to collect from as many individuals as possible, and to create an experience or mechanism that further retains them as long as possible to do as wide a range of things as possible. They are also incentivised to disguise the extent of this delegation, for example by embedding control protocols into complex and legalistic privacy policies whose acceptance is virtually costless (clicking the 'accept' button), binary (yes/no forever) and unconditional, and which are subject to arbitrary change without notice. Platforms get users to disclose as much of their information as possible (to the platforms' benefits) by artificially forcing a tradeoff between participation and privacy; in order to enjoy the most basic features of the Web, users have to *give their data away*, thereby sacrificing control over their data and potentially their privacy.

This misalignment of incentives between *what users want to do with their data* and *what platform providers want to do with their data* has the potential to interfere destructively with development of context-sensitive applications that promise more effective, personalised, behaviourally-adaptive interactions that rely on richer and more sensitive data models, due to either actual or perceived privacy risks entailed. Moreover, the dependency relationships that result from this process place unprecedented power in the hands of these companies, leaving individuals effectively locked in, and unable to switch to alternative providers without greater effort than it is reasonable to expect a privacy-aware non-technical consumer to devote to the problem; the result of this is an overall reduction of autonomy and mobility, potentially ultimately leading to increased fragility, fragmented data spaces and lost or forgotten data [**?**, **?**].

---

[1] Facebook Home - www.facebook.com/home

[2] Skydrive - www.microsoft.com/skydrive

[3] iCloud - www.apple.com/icloud

While this business model has thus far been hugely successful at creating extremely profitable services from the likes of Facebook, Twitter and Google, the result has been an increasingly fragile ecosystem in which a majority of Web users have come to rely on an oligarchy of service platforms which are in turn amassing a disproportionate quantity of users' personal information. This centralisation, and accompanying power asymmetry, has occurred not just for Web users from the United States, where most of these services are based, but internationally as well, raising concerns pertaining to each country's sovereign rights of access to data of its own versus other nations' citizens, which have been magnified by the information-gathering practices of the US National Security Agency and others revealed by Edward Snowden in 2013. Indeed, thorny issues pertaining to compliance and enforcement of data protection laws across international boundaries [**?**, **?**] represent a serious potential risk for this business model, even as the European Union debates a revision to its pre-Web Data Protection Directive. The EU's weak and unsatisfactory 'safe harbor' rule, which allows data sharing with the United States, conveniently diverting attention away from the unsolved problem of differing approaches to privacy and data, looks especially vulnerable — yet where would the cloud be without safe harbor?

However, a basic assumption that powers these dependence relations and underpins the oligarchy is the disparity between the data management capabilities held by the end-users of the Web from those that provide the hosting and storage. In this chapter, we question this "thin client" model of Web computing by examining an alternative approach that places the responsibility of data management back with the users who own it, but in a way that is natural and manageable, while supporting the same social, dynamic interaction flows they are used to on the Web. This set of capabilities we refer to as *personal data stores* (PDSs), the technical goal of which is to augment user computing devices with secure data storage, hosting, and sharing capabilities which can be used to archive and manage valuable information longitudinally, as they interact with one another and third parties respectively.

Our aim in this chapter is to derive the requirements for personal needs for such a platform through insights from the field of Personal Information Management (PIM). To begin with, it is worth reviewing in more detail the dilemmas and asymmetries that current management of "big data" has created, across the public and private sectors, and why the individual is understandably at a loss. Although PDSs cannot conceivably solve or even address all these issues, we should keep them in mind in order to understand the extent to which it makes sense to include PDSs as part of a more equitable longer-term settlement. Second, we present a brief summary of existing platforms being used to manage personal information and their characteristics. The chapter concludes with a discussion of how these platforms may change the socio-economic landscape of the Web, and the ways personal data is shared, collected and handled.

2 The Dilemmas of the Data Economy

Although we would not hazard a guess as to who originated the phrase, we do know that data has been called "the new oil" on many occasions. Of course, the image is intended less as an indication of the deep issues at the core of the data economy, and more as a neat way of conveying excitement in a Powerpoint bullet. However, it *is* indicative, because it can be taken in various, not necessarily exclusive, ways. Oil is a source of great wealth. It is a key factor in many other production and transport processes. It is an essential lubricant. It needs to be mined (well, drilled to be precise) to produce value. It brings great riches to the small number of corporations big enough to exploit it. It raises exchange rates and therefore prices to the detriment of other industries. It has been known to impoverish those whose property is drilled, as elites cream off the main wealth with the help of rapacious corporations and corrupt government. It has, on occasion, led to revolution and the overthrow of *anciens regimes*.

Presumably not all these phenomena associated with the old variety are intended to be predicated of the "new oil." Yet we, as data subjects, presumably want to be sure that we get the good things and not the bad. It is anyway a misleading comparison, because data has properties that oil does not. Data is about people, and can be compromising. Data is generated by people, not by aeons-old trees and animals which have no issues of privacy or dignity. It is not a dwindling resource — we are far fromapproaching the time of "peak data." It is not a rival good — if I enjoy its use, that does not preclude your exploiting it at the same time. Data becomes valuable when aggregated across communities. Data is a covert way of financing content and services; if the service you receive is free, then you must be the product.

Data is connected to us by an umbilical relation — we generate it in all sorts of ways, and it is about us. We create and provide it; we leave trails of it; it is inferred about us. Yet the flip side of this is that it expresses things we find important (indeed, Luciano Floridi [**?**] argues that our data are an inalienable part of our identity). By providing a route for others to understand what we are, or what we have done, or where we are situated, it can threaten our privacy, or our dignity, or our autonomy, by diluting the privileged first-person access to our own experience. It creates the possibility of our being counted, measured, judged, steered or influenced without our knowledge by mysterious forces or organisations who may or may not have our best interests at heart. And if the data about us don't exist, we can be profiled, and treated as a standard member of a small demographic, whether this is accurate or not. Maybe this means we get more interesting advertisements — or maybe we will be treated as a potential terrorist and denied access to an aeroplane without adequate explanation.

Because of this, data is regulated. Data protection law is intended to strike a balance between the public good (which may include commercial benefits) of data use, reuse and sharing, and the private good of privacy and individual control. Data subjects have certain rights over their personal data — but not the rights of ownership. If I browse an online bookstore, then I have thereby created a load of data which

is of value to someone else. They have constructed a website, and therefore claim ownership rights of the trail — the data results from their investment. In case of dispute, they will cite my consent to their use of my data via some privacy policy that I probably never noticed. It may be argued that I benefit from the collection of this data, because it gives the bookstore sufficient evidence to suggest other books to me that may interest me (and we know, from Amazon's early experience, that a good recommendation algorithm will easily outperform a human recommender). If I were given ownership of my browsing data, then there would be no incentive for the online bookstore to collect it, so it wouldn't be collected, and no-one would benefit from it. If data subjects owned their personal data, then third parties wouldn't bother to collect it, and the data economy would remain a glint in Google's eye.

Data protection is not there to protect privacy; that is at best its secondary purpose. But worse, data protection was a concept developed for the world of standalone databases, not the connected, networked Web with which we are familiar. A tangled skein of legislation struggles to cope with the realities of the personal data economy. Trading personal data goes on at scales previously unimaginable. A user goes online, and literally dozens of organisations will be tracking his or her behaviour. There is talk that this will benefit the data subject, via better devices, better websites and better recommendations; the main 'benefit', arguably, is to become a better target for marketing. I sacrifice my privacy and aspects of my intimate identity for a better class of spam.

Some economists [**?**] have argued that the release of personal data is a good thing for wider society, as it reduces information asymmetries and enables capital and currency to be allocated in a more informed way. Hmmm, maybe in an ideal world. But arguably the data economy is functioning by ramping *up* the asymmetries — data-using organisations not only know much more about the use they are making of their data than I do, they now in many cases *know more about me* than I do. I, the poor data subject, am sitting at the bottom of so many data asymmetries that the idea that I am too informed for the public good is surely laughable.

Furthermore the concepts of data protection, so valid and timely when they were first introduced, seem at best quaint in 2014. Data minimisation is a great principle, but is it realistic in a world where five billion Google searches and half a billion tweets are generated every day, not to mention the colossal number of mobile phone locations that are logged? Can the use limitation principle be of value in a world where serendipitous reuse is the order of the day? In the data economy, after all, the primary use of data *is* its secondary use. Do notice and informed consent have any meaning in such a world? Do we want to be notified of everything, when our lives are becoming increasingly complex and the choices we already have to make are multiplying? What, when I click the 'yes' button, am I consenting *to*? Doesn't virtually everybody treat the 'yes' button as an opening to a new and exciting online experience, rather than a notification of the commencement of a complex business relationship which entails a certain amount of risk for the insouciant clicker? Can this truly be glossed as *informed*? One might as well say that the fox is giving informed consent to the hunt when he tunnels under its boundary fence in search of prey.

But how to react to this? We must surely admit the many benefits that data can bring to the subject. Understanding oneself is an important part of managing one's health or consumption. The benefits accrue not only to the individual, but also wider society. Effective public health, transport and crime management are facilitated by giant quantities of accurate micro-level data. So data sharing with government and businesses cannot be made too difficult to do.

One potential way forward is to move from the current model of data protection, based on regulating the collection of data, for a defined purpose, centred on the data controller, and governed by the consent of the individual. The 'footprint' of the data now stretches far beyond the immediate context and purpose, and regulating for the moment of collection looks anachronistic. A number of commentators, including advocates of 'big data' [**?**], argue that the time is ripe for a move from subject consent to user accountability. Such a model would regulate the uses of data, and would be centred on the subject who would be given a greater, and less binary, measure of control. For example, Novotny and Spiekermann argue for a three-tier information market, with key distinctions in terms of responsibilities and liabilities between data subjects, service providers, a second tier service space that provides essential support for the top level service relationship, and a tertiary space in which data from the top level relationship is reused on an open but restricted market [**?**].

What kind of control should the subject be granted? Ownership brings responsibilities as well as rights, and as we noted may mean that potentially valuable data would never be collected at all. Furthermore, many thinkers are nervous that the concept suggests that people's identities are basically property and commodities, with all the dehumanisation that implies. On the other hand, a human rights approach, for example based on Article 8 of the European Convention, is something of a blunt instrument, and the article is frustratingly vague as to what we actually should do in a particular context. Furthermore, Article 8 is in place and agreed across Europe and many other countries, yet our data is still plundered by the data barons.

In the remainder of this chapter, we will attempt to answer some of these very difficult questions, by tracing a particular idea through conceptual beginnings to a concrete architecture. The next section will consider the notion of personal data or personal information in more detail.

3 Doing Things With Personal Data

Despite the clear importance of the concept, "personal data" means different things to different disciplines and communities. In this section, we will consider these different views of data with a view to understanding what capabilities it could afford for subjects, given sufficient access and control. We begin by looking at some of the different definitions, then from the perspective of Personal Information Management look at some of the activities around data and data management, and complete the section by considering the technological support for such activities.

3.1 What Constitutes "Personal Data"? Legal and Operational Definitions

The standard way to conceive personal data is via its legal definition, based on data protection law. This conception has two advantages: first of all it is widely accepted and understood, and secondly it matches the legal liabilities that any PDS management system will need to confront and accept. Personal data, on this definition, is data relating to an identifiable individual. There are a number of issues and indeterminacies here — identifiable by whom? using what methods? in what context? — but these need not detain us here, except to note that they do not make things any easier. The legal definition has not really kept up with technical developments, and it is clear that the ability to identify a data subject is highly context-dependent [**?**]. 'Personal data' is the usual European term, but in the US it can be known as 'personal information' or personally identifiable information'.

There are strong sanctions against the misuse of personal data without the data subject's consent, but data sharing can still take place if the data controller de-identifies the data by removing identifiers from it or aggregating it (whereupon the new dataset is no longer personal data). There are many techniques for this [**?**], and there is also a major and unresolved debate [**?**], [**?**], [**?**] about whether de-identified data can be made re-identifiable by cross-referencing it with other datasets, using so-called 'jigsaw identification' methods. For instance, the information that a girl in a dataset is pregnant is not identifying, and therefore not personal data, but combined with the information that Mary Jones is the only girl in the dataset, clearly a possibly unwelcome inference can be drawn about the all-too-identifiable Mary Jones. In this chapter, we will not consider the issues raised by such technicalities in detail, except to note that (a) they impinge on data sharing practices and may impose complex liabilities that will be hard to predict, and (b) they can be side-stepped in many cases if the data is exposed by the data subject, who can therefore be assumed to have given consent for use of that personal data by others, given that he or she made the decision to share it in the first place. In the context of giving data subjects greater control over the data that is about them, this is clearly a vital factor to consider.

If we now move from the legal definition, and consider this latter context, an alternative understanding emerges of personal information as the information over which a person has some interest or control, in order to negotiate their environment or order their lives (so, distinct from data in which a person has a commercial interest only). This type of personal information or data is much more in tune with an intuitive understanding of what data means to *me*. And as one would expect, it would include a great deal of crossover with the legal definition of the data from which I am identifiable, but it is likely to include data of which I am the owner, but from which I could not be identified at all (e.g. photographs I have taken, from which it may even be possible that other people might be identifiable, and hence which might be personal data with respect to those people).

The uses to which such data may be put might be social or entertainment, or could be work-related, consumption-related, or administrative; it might also have no obvious immediate use, but be stored in case it should have value later on. The data may come from several sources: it could be self-generated, deliberately created, a by-product of other kinds of activity, shared with friends or colleagues, open data from the Web, or have been officially bought or licensed from the (legal) owner. Therefore the data in which a person has an interest will almost certainly be of various types of legal status. Personal information in this sense has been investigated by researchers in Personal Information Management (PIM), and we can draw on some of their insights.

The task of identifying all of the kinds of data a person might need to keep, manage and use is a complex and not easily scoped task. Researchers in PIM have derived various working definitions of *personal information* in order to effectively scope their field of study, and and have made progress towards potential functional classifications for kinds of personal information. One such classification by Jones et al. [**?**] is visible in Figure 1.

Category	Examples
1. Owned/controlled by me	e.g., Email, files on our computers
2. About me	e.g., my credit/medical history, web history
3. Directed towards me	e.g., phone calls, drop ins, adverts, popups
4. Sent (provided) by me	e.g., Emails, tweets, published reports
5. Experienced by me	e.g., Pages, papers, articles I've read
6. Relevant (useful) to me	e.g., Somewhere "out there" is the perfect vacation, house, job, life-long mate

Table 1 Jones's 6 Types of Personal Information [**?**]

Jones takes an approach that distinguishes among different kinds of information by how it relates to the individual in question; whether the individual experienced it, kept it, sent it, or received it, or whether this information refers to the individual or his or her activities. The categories *About me* and *Relevant to me* are controversial because these definitions do not require individuals to be aware of the existence of the information; it thus establishes a sphere that goes beyond the scope of information experienced by the user. We discuss the potential implications of including such information within the scope of PDSes in *attentional challenges*.

3.2 Activities Around Personal Information

Each person can access, use and manage information in many different ways in their everyday activities. Moreover, there is considerable variation among the ways that different people manage their information, as documented in studies of people's office and home information environments predating personal computers altogether [**?**]. As a result, it has been relatively difficult to come up with a single characterisa-

tion encompassing all of these activities; several classifications have been proposed. Returning to the PIM literature, Jones et al. propose a categorisation centering about a distinction between finding, keeping, and a set of "M-level activities", which encompasses managing and organising information archives (Figure 2) [**?**]. Whittaker et al's slightly different categorisation, meanwhile, simply identifies 3 classes: keeping, management, and what they call "exploitation", as follows:

Jones [**?**], Jones and Teevan [**?**]	Whittaker et al [**?**]
(Re-)Finding	
Keeping	Keeping
Meta-level activities (managing, maintaining)	Management
	Exploitation

Table 2 Jones and Teevan vs Whittaker's categories of PIM activities

Jones's classification introduces *finding* as a primary activity that people perform; his definition spans a set of common behaviours including discovery [**?**], information foraging[**?**], orienteering[**?**, **?**], searching[**?**] among other related behaviours in which people purposefully seek information or serendipitously encounter it in the course of other information activities. Once this information is found, information is either consumed and internalised, or kept in an external archive, or both, and this process of saving information externally is referred to as *keeping*. Beyond this activity of archiving, individuals might return to their archives to organise, update, or trim them; such activities are referred to as the *M-level*, for manifold meta- and management, hence *M-level*, activities. Whittaker then includes a fourth category of behaviours, *exploitation*, referring to the the set of ways in which the information is used and applied.

Among such uses, while the foremost might be to *inform* an individual making a decision, many other uses of information also exist. For example, information might be created for the explicit purpose of *reminding* a person of past or future events, activities or details. Other purposes might be to *measure* and keep track of the time-evolution of some phenomenon so that it can be easily understood. When this measurement is about the individual's own activities, the purpose might be for providing *feedback*, which may be vital for behavioural modification domains such as cognitive behavioural therapy (CBT)-like programmes. This feedback may, in turn, along with other information, collectively serve to *motivate* further activity or behaviour. Finally, information may serve the purpose of *external cognition*, in which information is created or manipulated for the purpose of facilitating *understanding* or *problem solving*. This set of activities is often referred to as *sensemaking* [**?**].

3.3 Supporting Information Activities

Technological support for each of these information activities has demonstrated the potential to change not only how they are conducted, but the contexts in which they are applied. One salient example is that of Web search engines, originally created for Web page information retrieval, but which have become a nearly ubiquitous tool for accomplishing tasks across a much broader variety of activities, spanning both desktop and mobile. Another area is in supporting longitudinal curation; tools that automatically perform off-site, incremental, and continuous backup such as Apple's *Time Machine* [4] have become commonplace, allowing end-users to make their stored data more resilient to accidental deletion or data loss.

Yet technological support for most of the other aforementioned personal information activities, including reminding, sensemaking, discovery and orienteering, has remained rudimentary. Reminding in PIM tools, for example, has until only recently been limited to clock/calendar-based alarms that need to be explicitly set for a specific date and time, despite the rich variety of "off-line" strategies people have naturally adopted for their own uses[**?**]. While the basic calendar alarm remains heavily used, its precision, brittleness and intrusiveness have been documented to undermine effectiveness, sometimes through extended "snooze wars", in which users repeatedly dismiss alarms, resulting in their piling up over time. The alarm can end up a burdensome annoyance, instead of providing the intended assistance.

The mismatch between people's data management requirements and the technology to support it is not, of course, restricted to PIM. As another example where the promise has not been borne out, Privacy Enhancing Technologies (PETs) [**?**] have yet to make a mark either. They too have failed to transcend the perennial problem of demanding an investment of time and resources that few want to make, or want to have to make. They also put a relatively inflexible barrier between individuals and organisations, while the individual may in fact have very context-dependent requirements (it is handy, for example, for an online fashion company to know my size, even if I do not want this parameter value bruited abroad). Takeup has been predictably anaemic.

4 Personal Data Stores

Yet surely technology must be part of the solution to a technologically-driven problem. Technology creates data, with the connivance of the data subject, and tools have emerged for large-scale players to exploit their vast datastores. The concept we wish to explore in this paper, in response to the foregoing discussion of the challenges and context, is that of the Personal Data Store (PDS). This is a locus of control which leaves open a number of the key questions about ownership and property, while

[4] Time Machine - www.apple.com/uk/support/timemachine/

giving power to data subjects. Our aim in this chapter is to set out some of the possibilities of PDSs, and to try to show that at least some of the above dilemmas can be addressed with them. Clearly PDSs will not be the full story - but they should be part of the solution. We hope to suggest some ways this could happen, and how indeed it *has* happened, and to illuminate the potential by refining our account to produce a specific example of a PDS architecture.

The aim of PDSes is to start to narrow the aforementioned data inequality by bolstering the capabilities of individuals for managing, curating, sharing and using data themselves and for their own benefit. The idea is not for such capabilities to replace services, nor for individuals to take their data out of the rich ecosystems that exist today (a feat which would be practically impossible, not to mention potentially destructive), but instead to enable people to collect, maintain and effectively derive value from their own data collections directly on the device(s) under their control. The combination of such capabilities and derived value provides an incentive for individuals to take responsibility for, and invest effort in, the preservation and curation of their data collections, turning to external third parties for specialised services only where needed. The aim of such development would be to try to restore some balance by providing a locus for subject-centric management of data, to complement (and in some cases replace) the current paradigm of organisation-centric data management.

Arriving at an operational definition, we define PDSes as follows:

> A personal data store is a set of capabilities built into a software platform or service that allows an individual to manage and maintain his or her digital information, artefacts and assets, longitudinally and self-sufficiently, so it may be used practically when and where it can for the individual's benefit as perceived by the individual, and shared with others directly, without relying on external third parties.

This description leaves undefined the kinds of activities that might constitute "managing", "maintaining", "controlling fully" or "using" this information, nor even what kind(s) of information, owned by whom, that we are talking about. Fortunately, significant insight pertaining to many ways individuals readily use information (in both on-line and off-line contexts) has been gained through studies conducted at the intersection of psychology and computer science, particularly the Human-Computer Interaction (HCI) research community. Beyond insights about existing information practices, various ideas have been proposed dating back nearly a century about how technology might change human-information and human-human relationship, modulated by new emerging information technology.

4.1 Historical Reflections From Memex ...

The genesis of an individual-centric personal data archive pre-dates digital computers entirely, to Vannevar Bush's Memex vision of 1945[**?**], which proposed a mechanical framework for supporting the collection, archiving, and organisation

of information to facilitate later cross-reference and retrieval. Among the important contributions of this article was the significant emphasis on reducing the effort needed to capture and retrieve information, due effort being the primary impediment towards effective and frequent information use. To this end, Memex proposed that individuals could wear capture devices on their bodies (a camera strapped to the forehead), store such information compactly, conveniently and indefinitely, and retrieve it later through an associative mechanism modelled upon the human memory, queried naturally via gesture.

Two additional early projects that explored how such information archives might be realised were Ted Nelson's Xanadu [**?**] and Douglas Engelbart's NLS [**?**]. Both proposed that information environments could be interlinked through a global network of knowledge sharing, demonstrating many ideas in the 1960's that would not be realised in commercial systems for decades. While the former focused on hypertext and distributed collaboration, the latter focused on structured data collections, including data navigation, creation and management. Engelbart demonstrated an actual prototype of NLS in 1969, capable of synchronous collaboration, complete through a graphical user interface, that incorporated dynamic hierarchies, hyperlinks, and multi-view representations

The introduction of the personal computer (PC) in 1984 provoked the development of the first generation of digital personal information management tools, consisting of a variety of application software products designed to help individuals create and maintain collections of digital data, ranging from flexible, schema-agnostic personal database systems like Filemaker [5], to specific data types, such as digital calendaring tools, and "digital Rolodex" address books. Seeking to appeal to the first generation of personal computer users, many of these applications borrowed metaphors from paper-based information collection tools, from the notion of "documents", to that of files and folders, and even notebook ledgers and personal diaries. Along with this deliberate shaping of digital information into forms designed to be familiar with paper information organiser came interaction metaphors and organisation methods for them; from deletion of information by "throwing in the rubbish bin" to "desktop" and "filing cabinet"-based based information organisation and arrangement.

Meanwhile, research in personal information management continued to pursue the vision put forth by Memex, towards methods of automatically building archives of personal life activities and experiences, so that these might be used as external memory prostheses. The pursuit of this vision was partially responsible for the development of handheld and early wearable computing technology, such as the Xerox PARC Tab [**?**], arguably the first hand-held computer, which ran arguably the first automatic location-based personal lifelog, PEPYS[**?**]. Many systems that captured other aspects of context and activities soon followed, such as the Remembrance Agent by Rhodes et al., and the life archive by Clarkson et al., both at the MIT Media Lab's "Cyborg" Wearable Computing group. Since the breadth of kinds of activities and experiences that such systems captured transcended paper documents,

[5] Filemaker - www.filemaker.com

such research required re-thinking the shape of data away from paper-metaphors to other kinds of collections, including *information streams* (e.g., Lifestreams [**?**]) and chronological *lifelogs*, such as MyLifeBits [**?**].

The third, and potentially most profound, transformation of digital information tools occurred with Web 2.0, the rise of a "social Web" replete with dedicated apps and services for managing and sharing nearly any kind of previously imagined personal information, ranging from the sensitive and intimate to the public.

Meanwhile, the data proliferated too. Seeking to monetise the flood of information people were putting online, markets for personal information quickly began to emerge, prompting concerns over privacy, security, and rights of access, which in turn have driven government and regulators' interest towards giving citizens more protection over various aspects of how data about them could be collected and handled. This led to international efforts to craft data protection legislation, as discussed above. In terms of the provision of data to individuals, such legislation so far has focused on allowing data subjects to inspect the data an organisation holds about them; on receiving a subject access request, the organisation is obliged to correct inaccuracies, and to respect requirements that the data is not used in any way which may cause damage or distress, and that the data is not used for direct marketing purposes.

However, this is a fairly minimal power which is hardly congruent with the increasing clamour concerning rights to data, including the spread of enforced transparency of data from the private sector [**?**] and the vogue for freedom of public sector information [**?**], and technology (and technology policy) together with new attitudes to transparency bring more possibilities. In the UK, a government initiative called *midata* [**?**] is working to bring about the logical next step of customers getting direct and unfettered access to data kept about them by companies (other similar initiatives include the US Blue Button initiative[6] and the French Mesinfos group[7]). The ultimate success of *midata* will be contingent on several important steps in both technology and regulation, most particularly including realising effective tools such as personal data stores for letting individual users easily consume, consolidate and make use of this data once it is made available.

4.2 ... To Mydex: Birth of the PDS concept

Independent of such legislative approaches, both academic and industry-led efforts also began to commit resources to research towards identifying ways that end-user citizens might, in the face of the vast growing repositories of data being held about them, enjoy more control and privacy. An academic consortium known as *Vendor Relationship Management* (VRM) at Harvard's Berkman Center was realised to conduct multifaceted research into socio-legal-econo-technical approaches that might be employed. Among the products of this research was a vision that users

[6] www4.va.gov/bluebutton/.

[7] mesinfos.fing.org/.

might stand as their own information brokers, and start to act as peers with service providers, capable of negotiating fair and equitable mutual terms of data use during interactions with them[**?**]. Out of this work emerged the earliest mentions of Personal Data Stores for realising such capabilities in the context of online e-commerce, inspiring more than a dozen different Personal Data Store offerings, platforms and services backed by commercial start-ups since 2001[**?**].

As an example, consider Mydex, whose proof-of-concept offering dates back to 2009 [**?**]. Mydex designers worked with data-handling organisations to develop systems to support data transfer and sharing governed by consent and identity verification. Design principles included putting the individual PDS owner in sole charge of consent giving and revocation with a simple 'on/off' switch; giving the individual sole access to the private encryption key; verification of all organisations wishing access to data; and comprehensive data sharing agreements going beyond Data Protection Act protections. The business model for Mydex is still experimental, but currently the idea is to fund the stores by charging organisations for access to data; if the charge is set low enough, then they should save by side-stepping other access costs (e.g. the costs of writing a letter to the data subject). The Mydex services are currently free of charge to the individual. Mydex exploits cloud infrastructure with open source software, but its PDSs are discrete collections of files encrypted and controlled by the individual, including — and this seems prescient after the Snowden revelations[8] — the ability to choose the location of the data centre in which the PDS is stored. Similar open source personal data storage containers include The Locker Project[9], data.fm[10], Owncloud[11], and OpenStack[12], each of which provides various degrees of easy-to-set-up 'personal cloud' software that can be used to store and host content on the user's own server on the Web.

A consistent theme of commentary in this area has seen Personal Data Stores (PDS) as important, if not essential, capability for end-users towards growing a healthier global "personal data ecosystem". For example, an independent study commissoned by The World Economic Forum documented ways that the value of personal data might be further "unlocked", citing Personal Data Stores as a core enabling mechanism to turn end-users from consumers into more autonomous data brokers[**?**]. A separate comprehensive analysis by *Ctrl-Shift* on emerging commercial PDS platforms and offerings projected an enormous economic opportunity for PDS services in the next five years[**?**]. In their view, PDSs are the key to making sense of the myriad data sources that now surround us, from data we volunteer, to the data that commemorates observations of our behaviour, to the data inferred about us, combined with the data we generate via management of our personal affairs (e.g.

---

[8] www.theguardian.com/world/the-nsa-files, www.ub.uio.no/fag/informatikk-matematikk/informatikk/faglig/bibliografier/no21984.html.

[9] lockerproject.org.

[10] data.fm.

[11] owncloud.org.

[12] www.openstack.org.

in health or finance), and also bringing in data about our activities as customers or consumers, including our contributions to loyalty card schemes.

4.3 Failure to Launch: Barriers to PDS Adoption

Yet despite the extensive needs analysis and market potential identified, early personal data store offerings have thus far failed to attract substantial attention from users. While a number of factors are likely responsible, so the lack of interest among users has been attributed to the fact that many of initial PDS platforms have sought to simply re-create existing end-user experiences offered by popular apps and Web platforms, rather than creating new functionality. Despite the benefit that these PDS offerings provide in terms of data security, users are often less compelled to try something new if the tangible experience nothing new, while data security remains an abstract, inestimable threat which does not necessarily easily compel behaviour change [**?**]. Finally, since the very purpose of PDS offerings is to protect user data from third party access, these platforms cannot derive revenue from user data and must resort to subscription models — always less attractive to new users than than offerings that are completely free to use.

On top of these suppressors of the positive impulse to manage data, we must also remember that the markets work pretty well for some (the most powerful) operators, and so there is a great deal of inertia around. A dogmatic view of revealed preferences of course suggests that individuals' lack of interest in the technology shows they have no desire to curate their own data. They happily click on privacy policies they have never read, and they buy the goods that are marketed to them, at least in sufficient quantities to justify the marketers' costs. 'Push' models seem to be in the ascendant, because the data oligarchs are the only agents with access to the bigger picture of what data is held about you, what can be inferred from that data, what services are available, and how you relate to the general data context. 'Pull' models struggle, because individuals cannot see the opportunities that are around. In short, the argument is often made that the technological direction of travel is more or less set, that it serves the public good, that the public is uninterested in any alternative, and so, to coin a phrase, "get over it." This deterministic model has been called Zuckerbollocks [**?**], and it is important to challenge and resist it.

Heath et al write [**?**] that "there is market evidence that [the person-centric model of control over personal data] is starting to establish itself," but even they see a challenge to getting the model to work. Three conditions need to obtain simultaneously, on the account of Heath et al: PDSs must (i) make life simpler/better for the individual, (ii) appeal to data consumers by solving some of their problems (e.g. costs, or legal liability), and (iii) solve some pressing challenge that is holding back developers and entrepreneurs in this space. To these three, we can add a fourth, which is to rejig current data protection thinking. At the moment (2014), there are three key roles in the standard model of data protection: the data subject, the data controller and the data processor. The owner of a PDS is none of these (or none exclusively

— he or she is likely to be all three at various times), and it is hard to see how individuals can exercise autonomous control over the data that affects them without some recognition of them as active agents in a different kind of role. Furthermore, data protection legislation is intended to cover cases of personal data being misused by others; it does not cover cases where individuals accidentally (or deliberately) identify themselves. Of course, this is a reasonable starting point for protection, but if it is the only principle, it means that if an individual 'takes charge' of his or her data, he or she *loses* the cover of Data Protection Acts.

5 Six Not So Easy Pieces: Challenges towards Realising the PDS Vision

The goal of providing individuals with the capacity to maintain their own information longitudinally imposes a number of challenges to supporting the kinds of information activities we have described. In particular, we see six broad categories of challenge to be met; the first, most fundamental of which pertains to effective *longitudinal keeping*. Enabling individuals to keep their data safely for a long time, while ensuring its continued accessibility and usefulness impacts both the data formats and methods used to store them. For example, since a person's physical computational hardware is likely to fail with age, methods need to be in place for ensuring robustness to such failures, such as multi-device replication and easy migration from older to new devices over time. Moreover, as evidenced by Moore's law [**?**], since the technical capabilities and properties of such data storage devices and platforms are likely to change fundamentally, PDSes must be designed to accommodate (and take advantage of) such changes as they arise. The devices and technologies that have made the PDS vision possible date back only a couple of decades, whereas a safe haven for data such as we are envisaging might well have to last a working lifetime (before we even consider the issues surrounding inheritance of data after a death).

A second challenge is allowing individuals who might have little or no experience in the intricacies of data management to cope with the burden of data security and longitudinal maintenance. Using current tools and services, for example, managing your data yourself still means taking pains to ensure that one's personal data is not lost to hardware and software failure, malicious attacks, or safely migrated to new platforms and devices; such efforts require vast investments of time, effort and expertise. A general lack of expertise or willingness to do this means that people currently rarely know how, or bother, to back up or consolidate their data. Thus it is no surprise that individuals have been motivated to outsource maintenance of their data to third parties, such as cloud providers. In order to facilitate autonomy from such services, therefore, PDSes must seek to support directly, and automate where possible, tedious data maintenance tasks that have plagued PC users for decades. Such automation could both ensure compliance for promoting data security and integrity, such as continuous backup regimes, thereby countering recent studies of

the extremely low compliance of personal data backup and security maintenance practices [**?**, **?**].

A separate set of challenges arises from the shift back from service-provider controlled data storage to a user-centered model of data management. Although this will re-empower users to control the organisation of their data spaces, and eliminate the pervasive problem of data fragmentation [**?**], [**?**], the challenge with the increased flexibility that this approach affords is that it requires re-consideration of how third-party applications and services can interact with such data, which have traditionally been pre-defined to operate on a fixed, typically application-provider established, set of data representation(s) and manipulations. In a consolidated, user-centric data model, on the other hand, such representations may be be specified or modified by the individual, or by some other third-party application(s) on behalf of them, and thus applications themselves must be designed to accommodate such variability among representations.

The need to comply with local, national and international data handling requirements pose a fourth set of challenges. In particular, if PDSes are to support the storage of identifiable information, or more critically, regulated sensitive information such as individuals' medical records, then PDSes must implement a variety of security standards (e.g. []) to ensure secured storage. Perhaps more difficult might be achieving compliance with the additional data handling requirements imposed by these regulations beyond how it is stored and encrypted; in particular, key handling requirements and guaranteeing aspects of physical access to the machine(s). The integrity of data must also be secured — for instance, although a patient should have the right to challenge and correct inaccurate medical data, if the PDS is to store a version of medical data that is likely to be used (for example, in support of medical treatment in a foreign country), the data would need not only to be accurate, but also of appropriate provenance in order to be properly adapted to the standard workflows of medical treatment.

Even if PDSes were to achieve all of the aforementioned goals, individuals would still face the fact that service providers would inevitably continue to profile and amass information about them, as long as it aligned with their incentives to do so (and it is hard to imagine that it will not — for instance, a service provider may need to gather a large amount of personal data in order to ensure correct and appropriate billing for its services). Thus, if PDSes are to give users the degree of autonomy and independence from profiling, they would need to include privacy-enhancing technologies, such as IP anonymisers, user-agent randomisation and cookie blocking. This may be difficult or impossible to do on "closed" platforms such as iOS that prevent these techniques because they are perceived as "hacking".

Perhaps the ultimate set of challenges, however, pertain to accommodating change as it affects both the information itself and the practices and activities surrounding it, over the years that a PDSes is intended to operate. Technologies that bring in new ways that data is used and generated seem to be introduced every quarter, placing new demands how this information needs to be accessed, created and used. The most recent examples include wearable computing and "always on" wear-

able sensor technology, from simple devices such as Fitbits [13] and Fuelbands[14] that unobtrusively but nearly constantly measure simple aspects of an individual's activity, to complex computational devices that can both deliver and capture information in high fidelity and quantity anywhere, such as Google Glass[15]. Such devices, as well as innovative new apps in can in some cases bring about changes in norms pertaining to people's activities, including the ways people think about technologies themselves.

Looking forward at some of the ways such technologies might impact information activities, some have looked at the possible consequences and implications that ever-increasing information capture and access might have on the kinds of activities mentioned above. While Bell and Gemmel have argued [**?**] that such increased capture and access could create near-perfect records of our daily lives, allowing people to examine with unprecedented scrutiny their everyday activities, others such as Mayer-Schonberger have argued that such a utopian views overlooks a great number of potential unintended consequences [**?**].

The difficulties that this community has encountered have led us to reconsider, from the ground up, the need(s) these platforms are meant to address, so that they can be used to design a platform that will fulfill needs beyond secure data storage, towards new applications that promote the more effective use of data in both personal and social contexts.

6 Survey of Online Data Platforms and Services

Given this characterisation of the various kinds of *personal data* and activities around it, we can identify the ways that current online services fulfil the needs towards people's information types and activities.

Figure 3 characterises the top five personal data cloud platorms by number of users. While Facebook may not be considered and end-user personal data storage provider of the likes of Dropbox, it remains one of the world's largest brokers of personal information. Of particular interest is its introduction of Timeline in December 2011, when it started encouraging users to document the entire chronnology of their lives on the service, prompting users to backfill information about their lives from before they joined the platform through specific questions and prompts. As a result, Facebook has quickly amassed one of the world's largest single collections of lifetime biographical information directly elicited from individuals.

Facebook only the supports the storage of very specific information forms, spanning status updates, likes, photos, messages to individuals and so forth. While Google Apps and iCloud support similar structured data entries such as calendar entries, all but iCloud support general file storage. A survey of why people used

[13] Fitbits - www.fitbit.com

[14] Nike+ Fuelband - www.nike.com/fuelband

[15] Google Glass - www.google.com/glass

these storage services revealed that while backup had previously been the main reason for using online cloud services, multi-device access and sharing/collaboration have quickly eclipsed backup for reasons people use such services online [**?**]. The primary use of Facebook, meanwhile is to stay connected with others, as well as several emotional reasons, spanning reasons of self-actualisation and to fulfill the need to belong [**?**].

Facebook	Profile incl. Timeline; Friends; Events; Group memberships; Biographical history; States favourites; Preferences; Message archives; Liked pages, images, products; Places visited.	Free
Google Apps and GDrive	Any files; Google Docs; calendar; G+ profile; identify and profiles of friends; search history; page access history; bookmarks; locations visited	Freemium
Apple iCloud	iWork Documents, Photos, Calendars, Passwords (Keychain)	Freemium
Dropbox	Any files	Freemium
Skydrive	Office Documents; Any files.	Freemium

Table 3 Commercial third-party cloud storage offerings

However, these services primarily pertain to the management of a fraction of the personal data encompassed by Jones's definition above, specifically "data owned/controlled by me". If we also extend consideration to online services that host and collect "data about me" as well, there are now an increasing number of sensor-driven apps and services that facilitate the tracking of various, routine aspects of everyday life activities, spanning purchases, movements, wellbeing vital statistics; we list such life tracking sites in Figure **??**.

Service	Description	Logging Method
Foursquare	Visits made to points of interest	Manual check-ins
Moves	Complete history of a person's movements throughout the day as recorded from smartphone app	Sensed via smartphone app
Mint	Access to personal banking records (tracking spending)	Automatic
Withings; Runkeeper	Access to weight, blood pressure, heart rate	Semi-automatic
Fitbit; Fuelband; Jawbone	Daily activity levels	Sensed via worn sensor
Wattvision; Stepgreen	Energy consumption	Automatic (service provider)
Moodpanda; Mappiness; Gotafeeling	Mood	Experience Sampled
CalorieCounter; Fooducate	Daily calorie consumption	Manual

Table 4 Web lifelogging services that facilitate the capture and logging of everyday life experiences.

While both categories of services broker significant amounts of data, these do not generally meet the requirements for personal data stores, as service providers ultimately control how this data is stored, secured, and have full access to its contents. Other services, meanwhile have been launched ofcused on security of user data; a list of such services are listed in **??** and are sometimes referred to as the first generation of "personal data store" offerings.

Personal.com	Cloud svc for keeping important structured data of specific schema types (passwords, contact details)
Mydex	Cloud svc centered around specific structured data and identity verification

Table 5 Personal Data Store offerings which encrypt data to provide a high degree of user data security, e.g., only the user has access.

aerofs	Commercial solution for self-hosting a centralised dropbox-like service
bittorrent sync	Commercial peer to peer file synchronisation software for personal computers
gitannex	FOSS Distributed file metadata maintenance system for advanced users
cosicloud	FOSS self-hosted cloud platform for plug computers offering mail, photo, contact and metadata hosting and storage
data.fm	FOSS RDF-based Web data store with linked data support

Table 6 Self hosted personal data platforms

7 INDX: A Research Programme Around Personal Data Stores

The substantial challenges just described towards realising an actual PDS platform that achieves the goal set out in the introduction makes deriving a requirements specification daunting. Such a specification would require a well-defined and limited set of capabilities, provided in sufficient detail to be realised in a software (or software-hardware system). Yet, it is not clear how such a set of capabilities (out of many) should be chosen, nor how to choose a such a set to satisfy the requirement of minimality (to avoid overspecification). Nor, finally, is it entirely clear how to verify whether any such set could reach its intended goal.

Therefore we believe a research-centric, rather than development-driven, approach may be the most suitable for bridging the gap between the high-level challenges discussed and the evaluation of potential solutions. Towards this end, we have begun a research project centred about a set of core questions for investigation, and an open experimental research PDS platform called INDX[16].

[16] INDX source code and distributions - http://indx.es

The purpose of INDX and the research efforts around it, are several; from a research coordination perspective, it aims to serve as a common ground where various research communities may identify interrelated issues. This is a particularly critical role, as the kinds of work emerging from usable security, privacy, data durability, decentralised social systems, could both be informed by, and used to inform others about how approaches might fit into an integrated picture of future information management systems.

The second role is to serve as a base platform upon which various PDS technical and interface experiments can be tested in a real world setting. To this end, INDX will provide a basic implementation of what one might consider the most elementary kinds of services that PDSes are likely to need. We outline the specific such functionality in the next section. The reason that a complete, open implementation of a basic set of components is necessary for evaluation is to provide essential functionality to enable PDS researchers to focus on particular problems one at a time, rather than having to re-implement these basic components per experiment.

The third, and perhaps most critical reason for INDX is that a concrete implementation is necessary to even start to interrogate many of the goals pertaining to how the systems might be used by indivdiuals. A deployable implementation of a PDS architecture opens up the possibility of running field experiments, which can be vital to understanding how individuals might perceive or adopt functionality in actual use. Just as the social mechanisms of the Web could not be effectively studied until years after it was built (and continues to evolve), the various interface and interaction mechanisms of PDSes may set off different usage(s) that would altogether be difficult to anticipate prior to deployment. Such is particularly important for personal information management practices, which have been shown to be highly slippery and idiosyncratic; people appropriate and change the ways they use the tools in their collections in unexpected, creative ways in order to satisfy their particular needs.

7.1 Base functionality of the INDX PDS Platform

The base architecture of INDX consists of three components; a versioned database for semi-structured data, a distributed identity subsystem, and management logic that glues the components together. Each is described below, along with rationale for its design.

7.1.1 The Data Store

A key question in implementing the core component of a PDS is choosing the "right" database - what kind of data model should it use? What query language should it support? How should it store the data to ensure longevity?

As databases have evolved over the years, many kinds of database models have been proposed and improved. The INDX design process brought us to consider many popular database types, including "traditional" relational databases, document oriented (or "NoSQL") databases, graph based data stores, "XML" databases, and RDF triple stores, to name a few. Each offers a few distinct advantages over the others, and many open source implementations exist of each type.

Since there are several advantages to using pre-existing databases, the most obvious of which is the fact that using mature, open-source software is likely to be more reliable and require less engineering than creating a bespoke solution from the ground up. Beyond this purely practical development consideration, there is a greater argument for being database-agnostic [**?**], rather than sticking to a single implementation. In order to realise the PDS vision of longevity, an unavoidable fact is that hardware and software is going to change dramatically, as will the database systems built on top of them; moreover, there may be a need to accommodate a variety of different data demands, with uses and needs continually increasing, as data streams become more numerous, personal data archives archives become larger, and query and sharing functionality is tasked with increasingly challenging applications. What may make sense to run on a single "conventional" PC today might need to be run on a thousand nodes in some virtualised computer architecture in the future in order to accommodate an individual's increased storage and query capacity.

Therefore, using the age-old engineering principle of modularity, we sought to create the INDX PDS as an adapter on top of one or two basic underlying database systems. This decision has enabled us to target multiple databases at the outset, ranging from desktops and servers to mobile devices.

The question of finding an appropriate data and query model for PDSes is a more delicate question because the design choices made at this level are visible to, and thus directly affect, application developers, and to a certain extent, end-users. A variety of considerations need to be made when selecting the data model; first, whatever target model is chosen must be sufficiently flexible to accommodate (with reasonable transformation) the kinds of data that the platform will be managing. A poorly suited data model for the target will likely introduce inefficiencies that will either slow down performance, increase complexity or both.

Fortunately, most of the aforementioned data models are fairly general, each with specific characteristics; for example, relational databases require data to be factored into tables, which assumes a certain degree of data regularity; XML databases represent data as hierarchical structured documents; more general document-oriented stores manage collections of (either structured or unstructured) documents with limited metadata (comprising sets of keys for retrieval), while RDF ultimately represents data their granular components: triples.

Another dimension is that certain types of databases more typically afford guarantees that others do not; for example, many relational databases offer grades of ACID (Atomicity, Consistency, Isolation and Durability) guarantees [**?**], while few document-oriented or RDF triple stores do, partly due to technicalities arising from realising these guarantees in these settings. An additional advantage to relational databases is that extensive research on them has yielded well-known methods to

"tune" performance, such as ways to factor tables to avoid otherwise computationally expensive query operations, the creation of indexes and so on, whereas such methods and query performance predictability is remains less well established for other database types.

The culmination of these observations, with the availability of an highly respected implementation have led us to target a relational database, Postgres [**?**], for desktop and server hosted INDX stores.

7.1.2 Datastore Management

However, despite its large feature set, Postgres does not, "out of the box" meet all of the capabilities required of a PDS by the definition we arrived upon earlier. Given the need for PDSes to continue to meet changing information needs over an individual's lifetime, it is rather unlikely that any database will ever be devised at any point in time that will be able to fulfill all future information needs itself. Thus, this is where the design of the PDS has to provide incremental functionality extension, again, through encapsulation and modular design.

One of the immediate such functionality that must be added in order to use Postgres as the core data store is support for schemaless storage. Being a relational database, this is not straightforward; typical scenarios of the deployment of Postgres involves having a database programmer specifically create a bespoke set of schemas per data type being stored, consisting of tables and related views. Yet, in terms of PDSes, such needs may not be known at the time of set-up, and may change dramatically over time; moreover, it is practically impossible to know at design time the structure of all the data any user might want to store.

A second example also relevant to long-term data retention was providing the capability of a revisitable history of all data objects kept in the store. There are many uses for such a history, such as letting a user retrieve old versions of their objects, such as their documents, that were subsequently lost or altered, or determining how particular objects were changed over time. Such capabilities have started to become available in commodity software such as Apple's Time Machine, platforms such as Dropbox and Skydrive, as well as many collaborative software tools. Thus, we believe that it such a capability will soon become a standard capability assumed by users.

Other capabilities that in the works for INDX include managing replicated copies (for enhanced resilience against datastore failure and corruption), sharing (such as object-level sharing support), and encryption for handling sensitive data. Such platform-level data capability allows PDS platform application writers to take advantage of sophisticated functionality and data security without having to implement them within apps themselves, allowing the unilateral improvement of data handling without adding application-wise complexity.

An important piece of functionality that the PDS management logic also has to assume is to access control, which involves orchestration of at least three separate components: access control policies specified by the user and stored as rules, the

database's own gatekeeping mechanisms for granting access to the data kept within, and digital identities of users and applications requesting access, described next.

7.1.3 Distributed Identity Management

The current predominant model of identity management is that service providers perform this management directly for users; for example, service providers allow users to create principals with them, and provide authentication mechanisms as well. This model is inconvenient for a decentralised model of interaction, however, as it requires users to register a new principals with every single individual's PDS prior to interacting with them.

Distributed identity management protocols [**?**] offer a solution to this problem, by separating the problem of identity establishment and verification from its use. This permits, for example, an individual to grant access to sensitive data in their PDS to a verifiable identity of an entity, for example, their GP, even if ther GP has never previously interacted with their PDS. Currently popular distributed identity management implementations include OpenID[**?**], WebID[**?**], Mozilla Persona[**?**].

A related problem that is distinct to identity management is that of allowing third parties to request and securely receive access to data (with the user's permission). For this purpose, protocols such as OAuth[**?**] and SAML[**?**] have been developed and implemented across a large number of data providers, including Facebook, Instagram, and others. Such mechanisms allow these particular parties to continue to share data on behalf of the user once permission has been granted once, without subsequent user intervention.

INDX's reference implementation uses OAuth in conjunction with OpenID to allow interoperability with current Web services, particularly for the purpose of permitting transparent archiving of content that users distribute across the Web. It currently supports the archiving of content posted to social networking sites and services such as Twitter and Facebook, activity logging sites such as Nike+, Withings, and Moves, financial tracking sites such as Mint, open data sources such as OpenWeatherAPI, with support for other services to follow.

8 Looking Forward: Functionality for Future Information Management

In this final section, we wish to touch upon a few potential ways that PDSes might change the ways individuals will work with information in the future. A key goal will be to achieve consolidated data models from heterogeneous sources, for which we discuss the role of semantic technology and ontology matching and alignment algorithms; and the implications .

8.1 The Challenge of Automatic Consolidation

If one were to make an assumption that Personal Data Stores will eventually be able to draw in information obtained from hundreds to even thousands of third party data sources, for example, ranging from social networking posts to retail sites to Wikipedia to one's electronic medical record providers, so that such data may be safely archived, versioned and conveniently accessed, a question remains – how will this information be organised?

While this information could be kept separate and archived in its original form as provided, there are significant advantages to a user if this heterogeneous data is consolidated. By consolidation, we imply the act of combining complementary information from multiple sources into fewer, coherent and more complete and consistent representations. If this is done, like information items can be displayed in a consistent fashion, making coherent presentation and manipulation of items simpler; such cnsolidated information can be used by the user (and by the user's applications) uniformly, effectively eliminating the aforementioned problems of fragmentation mentioned earlier. The advantages to the user of a single consolidated data model are many, and we discuss a few of the potential ways this may enable applications to do more sophisticated things for users later in this section.

If all information service providers adopted the a single unified schema for all information coming into and out of them, this goal could be achieved relatively simply, since data records from separate sources could be directly compared. However, it is fairly well accepted that achieving such a singular data representation is as unlikely as convincing the entire world to speak exactly one dialect of a single language; the degree of diversity and continued independent evolution of systems practically guarantees that this will never happen[].

Thus to tackle this challenge, we must perform a kind of information integration, in which data are transformed into a consistent representation. For any pair of fixed, sources, bespoke mapping could be specified by a programmer manually. However, if the applications are not known, or if the data came in arbitrary forms unknown in advance (such as if they came driectly from a user), other methods must be employed. It is this latter situation that is likely to be quite common for PDSes, particularly considering the wide range of potential data and applications a user might need. We briefly discuss how semantic technology and ontology matching algorithms may be able to help.

8.1.1 Semantic technology

Research pertaining to the Semantic Web has looked at methods by which automatic inference over heterogeneous information can be made possible by grounding such representations in ontologies related through ontology languages, such as OWL[**?**]. Such semantics establish a framework by which machine translation of information representations become made possible through the formal stated connections made about such representations. The role of *semantic reasoners* thus are to take informa-

tion represented in such formats, along with their source ontologies, and to allow relationships among such information items to be deduced.

A requirement for such technology to work, however, is that all information providers provide appropriate mappings for their information representations against common ontologies using languages such as OWL. Thus far, few Web data sources outside of research and a few specific domains have embraced such techniques, making the use of such ontology languages, meaning that other approachese may also have to be employed. One such is the use of automatic ontology matching algorithms.

8.1.2 Ontology matching: automatic and interactive methods

Two other approaches have been taken to this problem; one is the use of machine-learning techniques for ontology matching (e.g. [**?**, **?**], or *instance matching* [**?**, **?**]). In such approaches, an algorithm is given a collection of examples of ontologies (or instances) and their corresponding semantic relationships, and the algorithm extrapolates properties to new, yet unseen relationships. This remains a rather computationally difficult task, however, and these methods have remained highly imperfect.

One promising approach has been to use such methods in combination with interactive approaches, that is to let users help such matching algorithms out when they get stuck. The *end-user programming* community has sought interfaces that can leverage information from non-expert individuals, who are empowered to assist and orchestrates the process of reconciliation at various levels of specificity. Systems that use this approach include "mash-up makers" (such as Mashmaker, [**?**], Marmite [**?**], Vegemite [**?**]) and interactive data workbenches, such as DataPalette [**?**].

8.2 Defragmentation and "Placeless" Data

One of the greatest advantages of the Web is that it has started enable pervasive information access; for an increasing proportion of the world's population, people can now access any information, anytime, anywhere from their desktops or mobile devices in nearly any setting [**?**]. Yet, the silos on the web have created artificial "places" in themselves; so now it is necessary "go to facebook" or "log into my university's portal" or "go to my health care provider", using the dedicated search and navigation facilities of these sites in order to get behind their *walled gardens* – even when the information being sought is the user's own information!

Such walls impede individuals' abilities to quickly access information needed, and in some cases, entirely preclude the ability for this information to be effectively cross-referenced, by preventing links from being established between these data items and increasing the barrier to accessing them. The result is often that the user experience of the Web has reverted back from the Memex vision of being able

to navigate fluid "association lines" of investigation aimed to complement the associative mechanisms of human memory and creative thinking, instead getting back to a series of online disparate bulletin board systems.

The vision of the PDS may reverse this at least for one's personal information, by providing consolidated representations of all of the information items distributed across silos that can be arbitrarily cross-referenced and linked. Doing this has its subtleties, however; as argued by, Marshall et al, "simply archiving" by harvesting a person's out of all of one's third party services necessarily decontextualises from the context of its original location, application or Web service in which it was created or found [**?**]. In order to avoid having this loss of context, PDSes could provide "wormholes" from the consolidated representation - which is better suited for sensemaking, to the individual serv ices hosting the rich context of content.

8.3 Supporting information management for life: Context-sensitive automation and behaviour change

Since it can be easily argued that the most valuable features of tools are the ones that are the greatest felt, we briefly touch on a few ways that the capabilities afforded by PDSes might directly impact people's lives.

The all too familiar feeling of data loss that occurs when we have had a hard drive fail, or the frustration that arises from not being able to find a particular important document or photo demonstrates the potential for technology to save people from distress in many immediate and direct ways. The position of personal information tools, as the most intimate and direct mechanism for satisfying a majority of our information needs, means that small changes in these tools can have substantial long-term effects.

Across many of the biggest information management problems are a host of well-known techniques that are simply not used because they are simply too time consuming, require expertise, or that people simply forget. For example, data loss can be practically avoided in relatively simple ways through the creation of off-site backups and vigilance in continuing to back data up over time. However, the low-compliance rate to backup regimens simply comes from the fact that people are often either too busy, forgetful, or simply do not know how to carry out such backups regularly. Similarly, limited time, attention, effort and expertise serve as the root cause for many other problems concerning long-term data preservation and access, including disorganisation, ensuring data security, and accidental deletion.

One potential solution to all such problems is the judicious application of automation in supporting a broader set of information management and maintenance practices. Just as spam filters transparently and automatically remove unwanted mail to save people from having to delete it themselves, or Apple's Time Machine continuously creates generational backups of the information on one's desktop and notebook without the user usually even being aware of it, we can imagine other in-

formation management activities being facilitated by more of such "attention free" support.

A particular kind of automation that has thus far been technically challenging to realise but well-suited to the capabilities afforded by PDSes is *context-aware* and adaptive automation that is sensitive to a user's needs, location, and activities. Since PDSes consolidate multiple information streams about a person's sensed activities (such as through wearable activity sensors or apps), it can consolidate the most complete digital "shadow" of the individual. This shadow, can, in turn be used by applications to provide attention-free automation support; for example, by using information about one credit card statement (such as from Mint) with one's current location (sensed via one's smartphone) and purchase history (collected in one's PDS over a long term), a future application might infer automatically that one is at risk of going over credit limit and intervene, either by warning the user, or automatically transfer money on his or her behalf to avoid over-transaction fees. Many such context aware scenarios have been proposed before (e.g. [**?**], [**?**], [**?**]), but their inability to get accurate, high-dimensional data of the user's context have impeded progress. PDSes seem an appropriate solution to this, particularly in situations such as the above where the involved in the inference is highly sensitive and personal, such as one's bank account balance, current location, medical conditions, and so on.

When such context-sensitive and adaptive approaches are applied to health and wellbeing, it can be used to play a role delivering better personalised coaching and intervention support. Simple forms of fitness coaching are already becoming available on the market, usually delivered as part of low-cost commercial activity sensors such as Nike's FuelBand, or Withing's body scale and blood pressure products. However, few of these applications are able to perform sophisticated tailoring due to the limited information available about the user from these single, simple sensor streams. Therefore, the kinds of multi-stream consolidation of user context may be helpful here towards more effective digital support in wellbeing maintenance, intervention and recovery.

9 Conclusion

In this chapter we have attempted to position the notion of Personal Data Stores as a (partial) response to the pressing problem of the autonomy of the data subject, and the asymmetry of power between the subject and large-scale service providers and data consumers. Given what Novotny and Spiekermann have called the "missing governance of personal data markets" [**?**] threatens to undermine subject trust in data sharing practice, and given that data sharing underlies not only a series of very valuable public services but also a whole economy, PDSs are highly suggestive of a means of putting the data subject at the centre of the data market's institutional structure.

The notion of 'personal data' is, for obvious reasons, in thrall to a legal definition that governs liability and policy, but the narrow legalistic coverage that this

subtends should surely be supplemented by a more intuitive notion of the data which is of interest, importance or value to individuals. Such a rethink would help both individuals, many of whom are concerned, if only in the abstract, that their privacy is being undermined by the collection, storage, aggregation and mining of their data, and data consuming organisations, many of which are concerned about a potential backlash. The rules governing ownership of data seem unlikely to change, as this would hamper the development of an equitable data economy, but regulatory and technical models are emerging in which the rights and responsibilities of various stakeholders are redistributed. PDSs are part of that emerging picture. It is also worth pointing out, however, that even with an unchanged regulatory position, PDSs have made some progress (e.g. [**?**]) — and the regulatory position is unlikely to remain unchanged in the charged atmosphere (at the time of writing) caused by the Snowden revelations and the revisions to the EU's Data Protection Directive.

Earlier, we set out six challenges facing PDSs, and described a reference implementation called INDX. The intention for INDX was not to make a claim that it would in its current state (or ever) solve all of the challenges, but to serve as a common artefact around collaborative research discourse for investigating sociotechnical issues and user needs. As a functional open platform, our hope is that it might be adopted as an instrument that accelerates research towards more flexible and adaptive information environments that assume dramatically different forms and shapes than our current models of silo-encapsulated hegemonies in the cloud.