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

No Author Given

No Institute Given

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 benefiting - 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 drive a reversal in such trends, 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. 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 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[**?**], Windows Skydrive[**?**] and Apple's iCloud[**?**]. Such services have extended the reach of Web services into the intimate digital spaces of one's personal digital 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 controlled, 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 such third party delegation is done in the context of the increasingly pervasive business model of deriving sustaining 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. 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) 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[**?**].

While this business model has thus far been hugely successful at creating extremely profitable services from the likes of Facebook, Twitter, Twitter and Google, the result has therefore 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 revealed by Edward Snowden in 2013. Indeed, issues pertaining to compliance and enforcement of data protection laws across international boundaries [**?**] represent a serious potential risk for this business model, as the European Union debates a revision to its pre-Web Data Protection Directive, which has been criticised for its weak 'safe harbor' rule which allows data sharing with the United States.

However, a basic assumption that powers these dependence relations and underpins the oligarchy is the disparity between the data management capabil-

ities 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 longitudinally archive and manage valuable information, as they interact with one another and third parties respectively.

Our aim in this chapter is to derive the requirements for personal needs for what 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 conceivable 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). It brings great riches to a small number of corporations that are big enough to exploit it. It raises exchange rates and therefore prices to the detriment of other industries. It has been known to bring impoverishment to those whose land is being drilled, as elites cream off the main wealth with the help of rapacious corporations and government departments. It has, on occasion, led to revolution and the overthrow of anciens regimes.

Presumably not all these phenomena associated with old oil are intended to be associated with 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. 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. We create it 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 (REF) 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 threatens our privacy, or our dignity, or simply to make our lives more complex than they would otherwise have been. 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 — the data results from their investment. In case of dispute, I will have consented 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, and it wouldn't be collected, so no-one would benefit from it. So 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.

So 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 Web with which we are familiar. The skein of legislation struggles to cope with the realities of the personal data economy. Trading personal data goes on at such a scale as to be almost uncontrollable. 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 (REF to Posner etc) 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 seem at best quaint. Data minimisation is a great principle, but is it realistic in a world where five billion Google searchs and half a billion tweets are generated every day? Can the use limitation principle be of value in a world where serendipitous reuse is the order of the day? In the data economy, the primary use of data *is* its secondary use.

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. Effective public health, transport and crime management require giant quantities of accurate micro-level data. So data sharing with government and businesses cannot be made too difficult to do.

If data subjects are to control data, what kind of control should this be? 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 implies that people's identities are basically properties and commodities, with all the dehumanisation that implies. Human rights, for example based on Article 8 of the European Convention, are 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.

Maybe technology can help. Yet Privacy Enhancing Technologies (PETs) have yet to make a mark. There is the perennial problem of their demanding an investment of time and resources that few want to make, or want to have to make. They also put a barrier between individuals and organisations, which the individual may not in fact object to (it is handy, for example, for an online fashion company to know my size).

The concept we wish to explore in this paper is that of the Personal Data Store. This is a locus of control which leaves open a number of the key questions about, while 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 to illustrate these with a specific example of a PDS architecture. To begin with, we consider the rich history of PDSs and related concepts into which we can tap.

3 Personal Data Stores: Introduction and Brief History

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 their own device(s) that they 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 individuals to manage and maintain their digital information, artefacts and assets, longitudinally and self-sufficiently, so it may be used practically when and where it is needed, and shared with others directly, without relying on external third parties.

This description leaves undefined the kinds of activities that might constitute "manage", "maintain", "control fully" or "use" 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.

3.1 Historical Reflections From Memex to Mydex

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 and conveniently

indefinitely, and retrieve it later through an associative mechanism modelled upon the human memory, queried naturally via gesture.

Douglas Engelbart's NLS[**?**] in 1969 demonstrated many ideas that would not be realised in any commercially available products for the next decade, including one of the earliest graphical user interfaces, the computer mouse, drag and drop manipulation, dynamic hierarchies, hyperlinks, hotkeys multi-view representations, and real-time remote collaboration. Finally, the introduction of the personal computer in 1984 provoked the development of many first generation personal information management tools, ranging from personal database systems like Filemaker [**?**], to digital calendaring and contact management tools, to file managers, spreadsheets and word processors.

Computer science research in the 1990s investigated approaches of automatic sensing and capturing aspects of everyday life activities into personal [**?**], such as for example the Pepys Memory Prosthesis [**?**]. Wearable and ubiquitous computing research continued this line of investigation, pursuing method of capturing of higher-resolution and more complete logs of people's activities (e.g,. MyLifeBits [**?**]), and applications for data-mining lifelogs for various important life patterns (e.g. Life Patterns [**?**]). The next decade saw specific evaluations of lifelogging in various specialised contexts, including healthcare for chronic disease maintenance, including memory prosthesis applications for Alzheimer's patients [], and cognitive behavioural therapy.

Simultaneously, the rapid spread of the Web brought a variety of apps and services for managing many kinds of information, ranging from the personal and sensitive to social to public. With increasing quantities of the population "going online" a market emerged for the personal information people were putting online, along with concerns over privacy, security over one's personal data, and rights to access. Government initiative to give consumers more protection over various aspects of both how data about them could be collected and handled were proposed and trialed with modest success in the United States and and more success in Europe. Simultaneously, independent research efforts in trying to give end users as consumers more control over their online privacy began to emerge such as *Vendor Relationship Management (VRM)*, which sought to not only investigate technical solutions but legal and economic frameworks that would lead to more beneficial outcomes for both consumers and businesses through consumer-empowerment [**?**]. Out of this work emerged the earliest mentions of Personal Data Stores, in the context of online e-commerce, which sparked from around 2011 more than a dozen different Personal Data Store offerings, platforms and services backed by commercial start-ups [**?**].

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 — the ability to choose the location of the data centre in which the PDS is stored.

The potential impact of personal data store technology towards driving new models of e-commerce and new experiences for end-users has been the focus of substantial interest recently among independent research organisations. The World Economic Forum commissioned a report on the personal data economy and ways to "unlock its value", outlining a programme projecting Personal Data Stores to be a core enabling mechanism through which emerging personal-data rich applications could thrive while simultaneously respecting the privacy requirements of individuals online[**?**]. Similarly, independent research organisation Ctrl-Shift also led a comprehensive analysis on emerging Personal Data Store efforts and their roles in information markets from a socio-legal-technical persepctive [**?**]. 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. in health or finance), and also bringing in data about our activities as customers or consumers, including our contributions to loyalty card schemes.

Data protection legislation allows 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. 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 [] and the French Mesinfos group []). 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.

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 for users. 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 - which are 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 simultaneous 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 it does mean that if an individual 'takes charge of his or her data, then he or she *loses* the cover of Data Protection Acts.

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 securely storing data , towards new applications that promote the more effective

use of them in both personal and social contexts. We first seek to establish a clear definition for PDSes based on a characterisation of what they were meant to achieve. Second, we derive a requirements analysis based on the abstract definition, deriving insights from the personal information management (PIM) research community.

3.2 What Constitutes "Personal Data? Legal and Operational Definitions

"Personal data means different things to different disciplines and communities. The standard way to conceive it is via its legal definition, which is 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 subjects consent, but data sharing can still take place if the data controller de-identifies the data by removing identifiers from it or aggregating it. 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 the 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 Mary Jones. In this chapter, we will not consider the issues raised by such technicalities in detail, except to note that (a) they do impinge on any 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 within a PDS only identifies its owner, who can therefore be assumed to have given consent for use of personal data with others, given that he or she made the decision to share it in the first place.

In another sense, personal information might be understood 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 other 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. If we focus only on relevance, the data may come from several sources: it could be self-generated, deliberately created, a by-product of other kinds of activity, be acquired by a midata-like initiative, shared with friends or colleagues, open data from the Web, or have been officially bought or licensed. It may or may not be owned by the PDS owner, though we must assume that the PDS owner has some control or rights over the data. Therefore legally, a PDS must be expected to hold data of various types of legal status. The personal information in this sense has been investigated by researchers in Personal Information Management, 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. from [**?**] is visible in Figure1.

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 Ive read
6. Relevant (useful) to me	e.g., Somewhere "out there" is the perfect vacation, house, job, lifelong mate

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

Jones takes an approach that distingushes 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*.

3.3 Activities Around Personal Information

Each person can access, use and manage information in many different ways throughout 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 for nearly a half century [**?**]. As a result, it has been relatively difficult to come up with a single characterisation 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 he called "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 bddoth, 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 them, update them, remove entries that have become unnecessary, and so forth; such activities are referred to as the *M-level* or *Management* activities above. Whittaker then includes a fourth behaviour, *exploitation* which he uses to abstractly refer to the ways in which the information is then used.

Among such uses, while the foremost might be to *inform* an individual towards 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 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* [**?**].

4 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, a tool originally created for Web page information retrieval, but which has 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 keeping behaviours; tools that automatically perform off-site, incremental, and continuous backup such as Apple's *Time Machine* [1] have become commonplace, allowing end-users to make their stored data more resilient to accidental deletion or data loss.

Yet such technological support has remained rudimentary towards most of the other aforementioned personal information activities, including reminding, sensemaking, discovery and orienteering. 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 result in their loss of effectiveness, sometimes through extended "snooze wars", in which users repeatedly dismiss alarms, resulting in their piling up over time. Such alarms in such cases end up a burdensome annoyance, instead of the assistance they were intended to provide.

5 Challenges Distinct to PDSes

The goal of providing individuals with the capacity to longitudinally maintain their own personal information imposes a number of unique challenges towards supporting the kinds of information activities just described. In particular, are four uniqu sets of challenges that must be met; the first, most fundamental of which pertains to effective *longitudinal keeping*. Enabling individuals to store safely keep their data for a long time, while ensuring its continued accessibility and usefuleness 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 the ability for data to be easily migrated 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 fundamentally change over time, PDSes must be designed to be able to accomodate (and take advantage of) such changes as they arise.

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

A second challenge concerns 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 with backing up or consolidating their data. Thus it is no surprise that individuals have been motivated to outsource maintenance of their data to third parties, such as cloud data providers we describe later. In order to facilitate autonomy from such services, thus, PDSes must seek to directly support, 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

A separate set of challenges arise from the shift back from service-provider controlled data storage to a user-centered model of data management. As mentioned earlier, this transition will re-empower users to control the organisation of their data spaces, rather than having it dictated by third parties. A second advantage to this approach is that it may eliminate the pervasive problem of data fragmentation [**?**], by allowing individuals to keep consolidated, definitive copies of their information, instead of being required to distributing information among separate services by their types [**?**]. However, 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. We discuss how semantic data representations have been used to address such challenges later in approaches.

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" wearable sensor technology, from simple devices such as Fitbits [2] and Fuelbands[**?**] 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

[2] Fitbits - www.fitbit.com

Glass[3]. 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 Blanchette have argued that such a utopian views overlooks a great number of potential consequences other factors [**?**].

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 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 [**?**].

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 **??**.

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

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

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 Technical Approaches

7.1 Proactive support: Context-sensitivity and automation

Location based reminding apps, such as Checkmark [4] have started to break out of aforementioned calendar-alarm to more adaptive reminding, by allowing alarms to be set sensitive to the user's automatically-sensed physical location. While still very simple and potentially fallible, this approach highlights the potential for greater context-sensitive support for all of the various personal information activities.

Context-sensitive support is particularly attractive towards future PDS work as it seeks to apply automation to provide

[4] Checkmark - itunes.apple.com/us/app/id524873453?mt=8