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 dynamically form new connections with others as needed, promoting a more adaptive, flexible and robust social fabric than those of traditional structures, in which efficiency lead a majority to rely on a handful of central, fixed intermediaries.

While we see technology being applied in many ways to support the kind of autonomy thus described, personal information environments 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 archives, trying to accumulate as many users as possible before the others. This trend as 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; content that started with a few social network profiles and blog posts, but gradually grew to encompass the entirety of personal data people keep, from files and documents to film and music archives. Thus began a migration of personal digital artefacts off the individually-administered personal computers into various information spaces of the web. The assimilation of personal data off personal digital devices has accelerated particularly recently, as Web application and service providers have started to create deep integrations with personal computing devices with examples such as Facebook Home[**?**], Windows Skydrive[**?**] and Apple's iCloud[**?**], respectively. 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 of managing 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.

While this mechanism has thus far been hugely successful at creating extremely profitable services of the likes of Facebook, Twitter, Twitter and Google, the result has been an increasingly fragile ecosystem in which a majority of Web users have come to rely on a handful of service platforms, which are, in turn amassing a disproportionate quantity of users' personal information. This centralisation has occurring 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, as well as issues pertaining to compliance and enforcement of data protection laws across international boundaries [**?**]. Moreover, the fact that these platforms are incentivised to get users to disclose as much of their information as possible has led to an artificial forced tradeoff between participation and privacy; in order to enjoy the most basic features of the Web, they 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 destructively interfere with development of context-sensitive applications that promise more effective, personalised, behavioural-adaptive applications 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 places unprecedented power in the hands of these companies, leaving individuals fundamentally powerless towards effectively switching to alternative providers in the long term; 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[**?**].

However, a basic assumption that powers this dependence 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* (PDSes), 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.

To derive the requirements for personal needs for what such a platform through insights from the field of Personal Information Management (PIM). 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 Background and Brief History

The genesis of digital personal data archives actually pre-date the digital computer 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. Douglas Engelbart's realisation of 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 was shortly followed by a many first generation personal information management tools for them, 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 [**?**], starting, perhaps with 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 rise of the Web brought an 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" emerged a market for the personal information people were putting online, along with concerns over privacy, security over one's personal data, and rights to access. Government initatives 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 the *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 [**?**].

The potential impact of personal data store technology towards driving new models of e-commerce and new experineces 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 [**?**]. Complementing this in the UK was a government initiative called *midata* [**?**] to give their customers direct and unfettered access to data kept about them by companies. The success of *midata* has been described to be contingent on several important steps, including realising effective tools such as personal data stores for letting indivdiual 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 exisiting 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.

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.

2.1 (Re-)Defining the Role of Personal Data Stores

The goal of the personal data store fundamentally to give individuals ability to safely keep, and effectively use any of their data for as long as they need it, and to share their data as they wish with whom they wish. Thus, we propose the following definition:

> A personal data store is a platform or service that allows individuals to manage and maintain their digital information, artefacts and assets, longitudinally, fully, and self-sufficiently, so it may be used practically when and where it is needed, 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 we are talking about. In order to approach a requirements analysis, one must consider both questions *what* and *how*; the kind(s), representation(s) of information to be stored and managed, and how the system is to support the user towards doing supporting use of and management of the data. Toward this end, the fields of information science and human computer interaction (HCI), particularly the research field of Personal Information Management (PIM), have worked to document the ways individuals work with, and manage the information in their lives, both in personal and work contexts. We thus propose that work on PDSes should be informed thoroughly by this literature, specifically in scoping *what* PDSes might do and further *how* they best do it.

2.2 What Constitutes "Personal Data?"

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*.

2.3 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) [**?**]. Jones 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* [**?**].

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, 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 transformational technological support has remained non-existent or at best rudimental for many 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 that have been documented in routine activities [**?**]. While such simple functionality is heavily used by many kinds of people, its precision, brittleness and intrusiveness have been documented to result in their loss of effectiveness, sometimes through extended "snooze wars" in which users reported repeatedly dismiss alarms to postpone them, resulting in them piling up over time, becoming a burdensome annoyance rather than help. Location based reminding apps, such as Checkmark [2] have started to break out of this model to more adaptive reminding, by allowing alarms to be set sensitive to the user's sensed physical location. While experiments towards more adaptive reminding strategies (e.g., with task based-reminding [**?**]) have been demonstrated in a research setting, they have yet to be realised in a sufficiently robust context for widespread use.

In discussing the support for PIM activities in the longitudinal context of PDSes,

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

4 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 [**?**].

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

4.1 Towards a Requirements Analysis

The next challenge, thus, is to identify how the design of future PDSes might be informed by such studies and abstract characterisations of the kinds of information people manage and the ways they go about doing so.

This is challenging for a number of reasons; first, the above characterisations are considerably high-level. Although examples can be generally identified for each, it may be difficult to extrapolate from specific examples a set of scoped design recommendations.

A second problem arises from the goal of long-term data management; since information can remain valuable for years or even generations, a central purpose of PDSes is to support the keeping and use of data over such durations. While merely storing data for such timespans is challenging given the rapid pace of digital infrastructures, data formats and computing platforms, supporting its effective use over such a duration may be considerably more difficult, requiring consideration of how not only the information might change but its use as well.

5 INDX: A Prototype Personal Data Store

5.1 Architecture

5.2 Implementation

6 Discussion

6.1 Ownership and Value

6.2 Data Licensing and Rights Management

6.3 Time-resilient Data Formats

6.4 Data Literacy

7 Conclusion