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 The Shape of Personal Data

At a lower level of detail, then, one might ask about the data model, its shape and, finally, its representation. For this purpose, Jones introduces the notion of *information forms* defined in terms of the tools used to manage them; e-mail is a single form because it is almost always accessed through an e-mail client, files though file management tools, blog entries using content management services (CMS) and so forth. While this approach may have worked well in the era when apps defined singular data types, over time, as tools have become more sophisticated, they have individually become able to handle more diverse and complex kinds of information. Perhaps more importantly, with the rise of Web-based tools, since all information is accessed through the singular tool of a Web browser, with data often spanning multiple sites and pages, such distinctions are difficult to draw.

A slightly different approach is to classify data by the data models and data schemas used. Since personal data can be derived from practically any application, they may be derived from fundamentally heterogeneous data models (e.g., network, relational, object-oriented, document, etc). Similarly, the source(s) may take any potential source schema. This means that PDSes must inherently address diversity at two levels, that of the fundamental data model, and that of the schema.

2.4 The Activities Around Personal Information

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

An additional requirement arises if PDSes are to present a unified, consolidated view of the user's data derived from multiple, heterogeneous data sources. Doing so requires addressing the challenge of reconciling heterogeneity. This is particularly critical as inconsistencies arising from data model heterogeneity can easily confuse users people and greatly increase the complexity of system use (e.g. [])

Finally, most data formats are not "future-proof" in the sense that they are not an open format that

3 Managing Personal Data Today: Services and Platforms

Today's data platforms are focused primarily around *backup* and *remote access*. The majority of cloud-based platforms offer both

personal.com	Cloud svc for keeping important structured data of specific schema types (passwords, contact details)	User keeps key, provider has no access to data	Subscription-b
mydex	Cloud svc centered around specific structured data and identity verification	Emphasis on identity assurance and certification	Subscription-o
Skydrive	Microsoft's cloud storage for personal	Freemium (5GB)	
Dropbox	Freemium (5GB)		
Google Drive	Freemium (7GB)		
Apple iCloud	Freemium (5GB)		

Table 3. Commercial third-party cloud storage offerings

Service	self-hosted	open-source	data types se
Dropbox	n	n	gf sp
bittorrent sync	y	n	
gitannex	y	y	n
cosicloud	y	y	sft/s n
SugarSync			
WD MyCloud			

Table 4. Self-hosted personal data storage platforms for end-users

While personal backup appliances such as ioSafe, WD My CLoud, and Apple's Time Machine have aimed at providing self-hosted data management encapsulated in simple "plug and play" data appliances for the home, adoption of even rudimentary home backup solutions remains low. Although current estimates are poor, it is thought that 10% of individuals effectively regularly back up their home computers, with the majority backing up only "periodically".

ioSafe
WD My Cloud
Time Machine

Table 5. Personal backup devices