7 Billion Home Telescopes: Observing Social Machines through Personal Data Stores

Max Van Kleek
Web and Internet Science
University of Southampton
Southampton, UK
emax@ecs.soton.ac.uk

Daniel Alexander Smith
Web and Internet Science
University of Southampton
Southampton, UK
ds@ecs.soton.ac.uk

Ramine Tinati
Web and Internet Science
University of Southampton
Southampton, UK
rt506@ecs.soton.ac.uk

Kieron O'Hara
Web and Internet Science
University of Southampton
Southampton, UK
kmo@ecs.soton.ac.uk

Wendy Hall
Web and Internet Science
University of Southampton
Southampton, UK
wh@ecs.soton.ac.uk

Nigel Shadbolt
Web and Internet Science
University of Southampton
Southampton, UK
nrs@ecs.soton.ac.uk

ABSTRACT

Web Observatories aim to develop techniques and methods to allow researchers to interrogate and answer questions about society through the multitudes of digital traces people now create. In this paper, we propose that a possible path towards surmounting the inevitable obstacle of personal privacy towards such a goal, is to keep data with individuals, under their own control, while enabling them to participate in Web Observatory-style analyses *in situ*. We discuss the kinds of applications such a global, distributed, linked network of Personal Web Observatories might have, a few of the many challenges that must be resolved towards realising such an architecture in practice, and finally, our work towards a fundamental reference building block of such a network.

Categories and Subject Descriptors

H.m [**Information Systems**]: Miscellaneous

Keywords

Web Observatories, Personal Data Stores, distributed systems, personal information environments

1. INTRODUCTION

The concept of a Web Observatory [30, 12] was introduced to investigate methods and mechanisms by which people, as a collective society, could be effectively studied in academic research settings, through the archival and analysis of the information traces they created online. As such traces have become increasingly rich, driven by both increased use of the Web and the onslaught of always-on smartphones, wearable sensors, and other devices that can measure the activities people perform on and off- the Web, two shifts have occurred. The first is that the boundaries between the activities previously considered "off-line" and those that were considered "online" is rapidly dissolving, meaning that all activities are being increasingly reflected in information about them on-line. A result of this is that the quantity, fidelity, sensitivity, and resulting value of this information is increasing - both in terms of potential value to individuals (as a multipurposable accurate record of their activities), and to third parties seeking to offer and provide services to people based on their lifestyle(s) and needs.

An implication of these two trends is that Web Observatories will no longer be solely about Web or what we currently think of as "Web-based activities" such as participating in online communities, social networks, and so on; rather, these observatories will be about the individual, multifaceted lives of people. From this perspective, it is unsurprising that significant privacy concerns may be raised about the large-scale collection of such data, whether they be for scientific study or commercial application. For example, even efforts driven by public bodies such as the NHS, such as the newly founded *Care.data* [1] have received widespread criticism (e.g. [21]) about its aggregation of millions of Britons' anonymised NHS patient records, even though such collection could drive medical research that might greatly advance the collective wellbeing [7].

This in general is reflective of a core dilemma faced in the building of such observatories; Web observatories will contain information of increasing potential value to making fundamental advances across research domains (spanning medicine, to human-centred design, to cultural anthropology, for example) but such repositories also represent unprecedented privacy risks and targets for identity thieves, misuse by commercial entities and so on, being comprised of aggregations of detailed, high-fidelity information about people's lives.

In this position paper, we examine one potential solution path towards resolving this dilemma: a technical architecture that changes the core assumptions surrounding the roles

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2578840.

[1]Care.data - A Modern Data Service for the NHS `http://care.data`

of data observer, aggregator, and broker proposed in Web Observatory research thus far. Specifically, we introduce the notion of *personal data store* as a core atomic component in a new kind of Web observatory; one that is purely distributed and in the collective control of all of its data sources – the individuals whose data form the observatory. We approach this idea by outlining the key functions of a Web Observatory, through what they are meant to achieve, which we follow with a definition of Personal Data Stores, including a summary of work done in this space before now. Then, we follow this up with the technical and societal implications of applying PDS architectures to building Web Observatories, focusing on identifying core challenges in this space, including preserving anonymity and privacy of members while promoting data sharing in such settings.

2. WHAT ARE WEB OBSERVATORIES?

A Web Observatory is a platform consisting of both a technical architecture and governance to enable the collection, sharing, querying, and analysis of Web Data [11, 30]. Given that the Web is a rich resource of the current state of the world, the aim of such Web observatories was set to provide a means to monitor, analyse and understand the activity of humans, both as individuals and as a collective. To do so, a core capability of such observatories is to combine information from many disparate streams of data generated by independent Web-based sources, spanning services, social network platforms, applications and so forth, into integrated coherent data models.

3. WHAT ARE PERSONAL DATA STORES?

The rise of "Web 2.0" was marked by transition of the Web from an information publishing medium to being a general platform for all sorts of human interaction, spanning from synchronous point-to-point interaction to many sorts of one-to-many information exchange mechanisms. As Web platforms became more sophisticated and complex, however, we also observed a trend towards greater centralisation; although many factors were involved, among them was the fact that building complex Web services and applications simply required more investment and expertise than most individuals could themselves muster; therefore, the construction of such services quickly became the domain of venture-backed startups. These startups, the nascent Facebooks, Dropboxes and Googles quickly amassed huge quantities of personal information as individuals flocked to their use for their services and capabilities. Seeking to derive revenue from such troves of user information, such companies forged the first versions of now a multi-billion pound a year surveillance-and-analytics business model. Although the kinds of content being amassed began as a few social network profiles and blog posts, it quickly 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 aim of PDSes is to start to re-balance the this data inequality by bolstering the capabilities of individuals for managing, curating, sharing and using data themselves and for their own benefit. The idea is not for such capabilities to replace services, nor for individuals to take their data out of the rich ecosystems that exist today (a feat which would be practically impossible, not to mention potentially destructive), but instead to enable people to collect, maintain and effectively derive value from their own data collections directly on the device(s) under their control. The combination of such capabilities and derived value provides an incentive for individuals to take responsibility for, and invest effort in, the preservation and curation of their data collections, turning to external third parties for specialised services only where needed. The aim of such development would be to try to restore some balance by providing a locus for subject-centric management of data, to complement (and in some cases replace) the current paradigm of organisation-centric data management.

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

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

This description leaves undefined the kinds of activities that might constitute "managing", "maintaining", "controlling fully" or "using" this information, nor even what kind(s) of information, owned by whom, that we are talking about. Fortunately, significant insight pertaining to many ways individuals store, archive and retrieve information, in both on-line and off-line contexts throughout the course of many every-day activities, has been the focus of a studies of the field of Personal Information Management (PIM) (e.g., [25, 3, 32]). Such studies have documented the breadth and often idiosyncratic nature of personal information practices, driven by both the fragmented nature of people's information spaces (arising, in part from the lack of integration among apps and siloed data sources of the Web), and the remarkable ingenuity with which individuals often worked around such limitations in order to manage their information archives. Since PIM studies have uncovered, in nearly equal parts, areas where digital information tools have served people well, and those where they have fallen fantastically short, this literature served as a convenient starting point for deriving the needs for PDSes. Our design process for our PDS, described later, thus started with a broad consolidation of results of these studies [31].

4. PERSONAL WEB OBSERVATORIES

Combining the two ideas of a PDS with the goals of a Web Observatory, a logical first step we propose is that of a *Personal Web Observatory* (PWO), a technical platform that, first and foremost, enables individuals to consolidate and archive their data currently dispersed among multiple sources. Then to use such a consolidated archive to serve as a kind of "analytical mirror" that can enable an individual to accurately gauge and reflect upon the multifacted states of their lives and wellbeing. Such consolidated data could be used, for example, for better time budgeting, stress management, budget-planning (through the consolidation of data streams representing spending), fitness and health manage-

ment (such as through sensed data streams representing the individual's vital statistics and activities).

With more sophisticated functionality, such a PWO might monitor one's social interactions, and correlate such interactions with states of wellbeing; do certain people seem to be the sources of stress or enjoyment? Similarly, such information might be used to 'debug' an individual's other states of wellbeing, such as to identify why random, sporadic headaches might be occurring, such as by correlating such incidences with particular activities, sleep levels, presence in particular locations, with certain times of the year or periods of the month, or with certain activities. Such "small data" analytics, while sparse, could be made statistically viable when gathered longitudinally over time, and offer the advantage that they reflect a single person's idiosyncratic patterns and correlations.

5. LINKED, DISTRIBUTED, PERSON WEB OBSERVATORIES

A next logical step from a PWO, then, is towards overcoming the single-individual limitations of a PWO by enabling individuals to combine their data. The remainder of this paper examines what such a capability would entail, and proposes our progress towards a potential implementation of such a system, a platform we call INDX.

5.1 Scenarios

Prior to identifying barriers to achieving such a goal, we first identify the kinds of *usage scenarios* we envision linking PWOs might enable.

5.1.1 Distributed QA

One broad class of uses can be thought of as a mixture of collaborative software and online Question-Answer sites [13], in which individuals can issue distributed queryies to a community for things that he/she needs to know. Like the now-defunct distributed QA service Aardvark [15], such queries might be cast to a specific set of people (such as acquaintances or members of an organisation, for example), or, they might be routed to the most qualified or available individuals. Unlike Aardvark, however, in which users answered all questions, in the PWO scenario, such queries might be posed in a machine-parseable form to allow individuals' PWOs to automatically service them.

Perhaps the most beneficial capability might come from the ability to aggregate responses to such queries automatically across an entire population. Such a capability could be used to allow individuals to gather large scale statistics useful for computing metrics such as bounds for realising differential privacy [8] policies. For example, it could be possible to ask specific demographic questions such as *"How many people live in Southampton who have a pet terrier, wear glasses and work at a Costa?"*. While seemingly oddly specific, the response to such a query might be used to determine the degree to which an individual who wishes to remain anonymous in public (to protect themselves, for example) might choose to be more selective about what they disclosed about their activities or employment.

However, for such scenarios to be realised, methods to ensure that such questions can be answered themselves without violating the privacy of the responders, question answers, or intermediaries will likely be essential.

5.1.2 Publishing Profiles for Data Analysis

While the previous scenario discussed a "pull" approach to distributed data analysis, another approach is to essentially allow groups of individuals to "push", or expose, "public profiles" of particular aspects of themselves for purposes that serve the collective good. For example, if smart automobiles in the future volunteered their coordinates (in a privacy preserving way) in real time to a collective tally of road congestion, people could in real time determine which routes to avoid, entirely obviating the need for a centralised service to do so (such as the Waze App[4]).

Just like in the previous scenario, an essential privacy requirement might be for such profiles to provide selective but authoritative statements about someone or something being in a particular state or having a particular property, without identifying the individual or thing that has it. Similarly, it would have to be guaranteed that multiple such profiles could not be attributed ot the same source, a well-known form of *disclosure intersection attack* [6].

5.1.3 Ethnographic Enquiry and Web Science

Beyond the specific sorts of data push and pull to support the kinds of queries and analysis described above, a third set of applications might be in supporting effective ethnographic enquiry and analysis in such environments, were individuals possess vast repositories of information about their daily activities and experiences. Answering such a question may necessarily involve addressing the issue of information legacy, and how one might support the effective preservation, ownership rights and control of life activity databases stored in people's PWO from one individual to the next, after they have died. The complex moral and culture-specific issues pertaining to addressing such problems have been discussed extensively elsewhere (e.g. [26, 9]) and are particularly salient for PWOs, where individuals might be in possession of complete records of their own life histories.

5.2 Towards Linked PWOs: Challenges

The goal of realising the previous visions of interlinked PWOs requires addressing a large host of challenges, from those pertaining to the PDS-level challenges of longitudinal information keeping, the many privacy-related challenges pertaining to effective information disclosure without privacy loss, dealing with attackers and identity thiefs. We outline a few such challenges we have not already discussed, below.

5.2.1 Long-term Data Maintenance

Enabling individuals to keep their data safely for a long time, while ensuring its continued accessibility and usefulness impacts both the data formats and methods used to store them. For example, since a person's physical computational hardware is likely to fail with age, methods need to be in place for ensuring robustness to such failures, such as multi-device replication and easy migration from older to new devices over time. Moreover, as evidenced by Moore's law [24], since the technical capabilities and properties of such data storage devices and platforms are likely to change fundamentally, PWOs must be designed to accommodate (and take advantage of) such changes as they arise.

5.2.2 The End-user Expertise Gap

A core philosophy of participant-centric PWOs is that the user assumes all responsibilities for managing and securing their data, as well as making critical decisions regarding their privacy, their own security, and ways to apply PWOs to their tasks and responsibilities. This saddles users with significant burdens which may both be extremely effort-expensive, but that individuals might actually have no expertise, experience or interest in doing. From this perspective, it is no surprise that, even in these comparatively simple days of "Web 2.0" data management services, individuals have been motivated to outsource maintenance of their data to third parties, such as cloud providers.

5.2.3 Third-party Interoperability

A separate set of challenges arises from the shift back from service-provider controlled data storage to a user-centred model of data management. Although this will re-empower users to control the organisation of their data spaces, and eliminate the pervasive problem of data fragmentation [17], [14], 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.

5.2.4 Handling Identifiable Information

When multiple individuals' PWOs interact and exchange data, the handling of others' data may constitute the handling and storage of *identifiable information*[20]. The handling of third-party identifiable information places, under many current forms of legislation, in a category which requires them to comply with local, national and international data handling requirements. Such requirements are more sever if some of the data exchanged fall in the category of particular kinds of sensitive information, such as individuals' medical records or histories, in which case PWOs must comply with a variety of stringent requirements (e.g. [1]) to ensure secured storage.

5.2.5 Anticipating Future Needs

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 PWO 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[3] that unobtrusively but nearly constantly measure simple aspects of an individual's activity, to complex computational devices that can both deliver and capture information in high fidelity and quantity anywhere, such as Google Glass[4]. 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 [2] that such increased capture and access could create near-perfect records of our daily lives, allowing people to examine with unprecedented scrutiny their everyday activities, others such as Mayer-Schonberger have argued that such a utopian views overlooks a great number of potential unintended consequences [19].

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 fulfil needs beyond secure data storage, towards new applications that promote the more effective use of data in both personal and social contexts.

6. INDX: A REFERENCE PWO "ATOM"

In this section, we briefly introduce our efforts at designing a reference implementation of first PWO Atom, an open source community platform called INDX[5]. While hoping to solve all of the aforementioned problems may seem foolhearty, the goal of our efforts are to try to identify, rather than solve in an ideal manner, existing methods and technology that can applied to make incremental progress towards various dimensions of an interlinked, global Personal Web Observatory. Embracing the philosophy of Richard Feynman ("What I cannot create I do not understand" [10]) we have found that the process of designing a PWO itself has surfaced both unanticipated challenges that could be solved with a practical application of (some still emerging) systems architecture best-practices. We briefly discuss design challenges pertaining to the following four areas: distributed sharing, authentication, synchronisation, longitudinal storage, and anonymous distributed querying.

6.1 Distributed architecture considerations

The problems of distributed sharing include issues of trusting external servers to be who they claim; and determining which information should be shared and which should be kept private. In distributed architectures, exchanging information might might involve both simple direct point-to-point communications, as well as communications relayed through any number of (potentially untrusted) parties. In either case, the ability for both communicating parties to establish a secure channel to one another with the guarantee that the other party is the intended one is essential for secure information exchange to be possible.

The problem of authentication is that of being able to verify the identity of entities, including users, within a distributed system. In a traditional system, a user would typically log in to each system explicitly to authenticate with it, typically first establishing a principal for each first. However, in a distributed system, explicitly establishing principals on every system is inefficient, requiring the creation and maintance of $O(n^2)$ principals. Distributed identity systems, such as OpenID [22], WebID[16], or Persona [18], discussed

[2]Fitbits - `www.fitbit.com`
[3]Nike+ Fuelband - `www.nike.com/fuelband`
[4]Google Glass - `www.google.com/glass`
[5]INDX: A personal data platform - `indx.es`

earlier, meanwhile provide a solution that uses a proof-of-identity mechanism relying on common third parties, which are typically well-known *distributed identity providers*. We have added OpenID support into the INDX reference implementation to allow users to prove their identities to any other INDX Atom, and support for other protocols is planned for the future.

Synchronisation refers to the ability to support concurrent editing of shared information items in a partially-disconnected environment, such as when an INDX node is occasionally powered off or when network connectivity becomes sometimes unavailable. Allowing shared information items to continue to be edited, even when some nodes where copies are stored are unavailable means that changes must be reconciled when communication among nodes is re-established. Methods have been devised to support user-intervention-free reconciliation such as [27], and INDX currently takes a simple, opportunistic approach to handle a majority of such cases without user intervention. Similarly, the challenge of durability against data loss, described earlier, is addressed in INDX through an implementation of the LOCKSS principle [23], in which important data is automatically replicated across several INDX instances, located on physically separate and potentially distant locations, to reduce the likelihood of data loss.

Finally, the INDX instances as PWO "atoms", for the kinds of applications envisaged earlier, requires consideration of both how such queries can be effectively performed, and ways that individual participants' identities can be effectively protected in the process. Although this remains an area of active research for INDX, we are drawing upon work in decentralised information indexing and query routing in peer networks (e.g. [5, 29] as well as methods that preserve anonymity by considering methods such the conceptually simple *k-anonymity* [28] to the theoretically-grounded methods of *differential-privacy* [8] for protecting participants' privacy under query disclosure. Specifically, we are considering these methods for allowing INDX users to easily express *how* identifiable they want to be, and then automatically deducing an appropriate exposure policy for answering distributed queries.

6.2 The Wellbeing Observatory

Our first PWO application for INDX is the Wellbeing Observatory, an application which aims to demonstrate ways that information fusion and distributed query can benefit an individual in a health and wellbeing context. The idea of of the Wellbeing Observatory is to consolidate information from the large number of worn activity sensors and devices that measure individuals' daily activities into a singular, coherent diary of their daily lives. Abstracting raw sensor signals to approximations of physiological signals also guarantees that, even as sensors are lost, replaced, or made obsolete, the information representation will remain consistent in terms of standard physiological concepts and measurements.

With respect to social PWO functionality, the observatory will offer an ''ask the crowd" feature which will allow individuals to ask others (with the option of doing so anonymously), which will route the question to an appropriate set of individuals who do not even need to be acquainted. For example, if an individual is trying to identify the cause of a particular set of symptoms they are experiencing, they might query the crowd for others with similar medical histories, living in their geographic region, or with similar recent activity histories to determine whether others have experienced the same symptoms and why. An early interface mockup of the functionality we envision in the wellbeing observatory is visible in Figure 1.

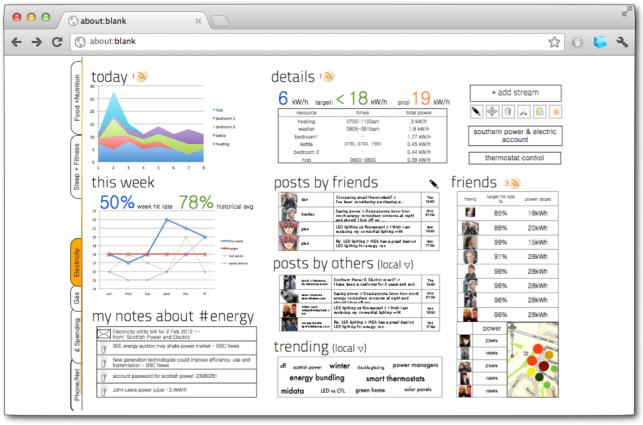

Figure 1: An interface mockup of the Wellbeing Observatory to be integrated into the INDX Personal Data Store platform; this application will showcase sensor data integration into a consolidated representation of an individual, and social querying functionality.

7. CONCLUSION

In this position paper, we proposed a technical approach to building a Web Observatories comprised of singular components centred around the individual. These interconnected components, which may be all different, will be based on the Web and exchange data and interoperate fluidly over time, even as the technologies they are based upon change beneath them.

Just as the most powerful radio telescopes are formed by thousands of smaller telescopes, arranged and connected in a way to form a coherent array more capable than any singular node, we feel that a billion-node linked PWO connecting every human's personal data on the planet might one day allow questions about humanity and society to be answered at depths and scales unreachable by any other means or instrument.

8. REFERENCES

[1] D. Banisar and S. Davies. Global trends in privacy protection: An international survey of privacy, data protection, and surveillance laws and developments. *John Marshall Journal of Computer & Information Law*, 18(1), 1999.

[2] C. G. Bell, J. Gemmell, and C. Rosson. *Total recall*. Dutton, 2010.

[3] M. Bernstein, M. Van Kleek, D. Karger, and M. Schraefel. Information scraps: How and why information eludes our personal information management tools. *ACM Transactions on Information Systems (TOIS)*, 26(4):24, 2008.

[4] A. J. Blatt. Technological changes in maps and cartography. *Journal of Map & Geography Libraries*, 9(3):361–367, 2013.

[5] P. Cudré-Mauroux, S. Agarwal, and K. Aberer. Gridvine: An infrastructure for peer information management. *IEEE Internet Computing*, 11(5):36–44, 2007.

[6] G. Danezis and A. Serjantov. Statistical disclosure or intersection attacks on anonymity systems. In *Information Hiding*, pages 293–308. Springer, 2005.

[7] S. de Lusignan and C. van Weel. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*, 23(2):253–263, 2006.

[8] C. Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[9] J. C. Fernández-Molina and E. Peis. The moral rights of authors in the age of digital information. *Journal of the American Society for Information Science and Technology*, 52(2):109–117, 2001.

[10] R. P. Feynman. Simulating physics with computers. *International journal of theoretical physics*, 21(6):467–488, 1982.

[11] W. Hall, T. Tiropanis, R. Tinati, P. Booth, and P. Gaskell. The Southampton University Web Observatory. In *Workshop on Building Web Observatories (BWOW) at the International Web Science 13 Conference (WS13)*, pages 1–4, 2013.

[12] W. Hall, T. Tiropanis, R. Tinati, P. Booth, P. Gaskell, J. Hare, and L. Carr. The Southampton University Web Observatory. pages 1–4.

[13] F. M. Harper, D. Moy, and J. A. Konstan. Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 759–768. ACM, 2009.

[14] T. Heath and C. Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[15] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on World wide web*, pages 431–440. ACM, 2010.

[16] G. Huang and K. Mak. Webid: a web-based framework to support early supplier involvement in new product development. *Robotics and Computer-Integrated Manufacturing*, 16(2):169–179, 2000.

[17] D. R. Karger and W. Jones. Data unification in personal information management. *Communications of the ACM*, 49(1):77–82, 2006.

[18] H. Koshutanski, M. Ion, and L. Telesca. Distributed identity management model for digital ecosystems. In *Emerging Security Information, Systems, and Technologies, 2007. SecureWare 2007. The International Conference on*, pages 132–138. IEEE, 2007.

[19] V. Mayer-Schönberger and K. Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray, 2013.

[20] A. Narayanan and V. Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 53(6):24–26, 2010.

[21] R. Ramesh. Nhs patient data to be made available for sale to drug and insurance firms. *The Guardian*, January 2014.

[22] D. Recordon and D. Reed. Openid 2.0: a platform for user-centric identity management. In *Proceedings of the second ACM workshop on Digital identity management*, pages 11–16. ACM, 2006.

[23] V. Reich and D. S. Rosenthal. Lockss: A permanent web publishing and access system. *D-Lib Magazine*, 7(6):14, 2001.

[24] R. R. Schaller. Moore's law: past, present and future. *Spectrum, IEEE*, 34(6):52–59, 1997.

[25] A. J. Sellen and R. H. Harper. *The myth of the paperless office*. MIT press, 2003.

[26] R. Shields. *Cultures of the Internet: Virtual spaces, real histories, living bodies*. Sage, 1996.

[27] C. Sun and C. Ellis. Operational transformation in real-time group editors: issues, algorithms, and achievements. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 59–68. ACM, 1998.

[28] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[29] I. Tatarinov, Z. Ives, J. Madhavan, A. Halevy, D. Suciu, N. Dalvi, X. L. Dong, Y. Kadiyska, G. Miklau, and P. Mork. The piazza peer data management project. *ACM Sigmod Record*, 32(3):47–52, 2003.

[30] T. Tiropanis, W. Hall, N. Shadbolt, D. De Roure, N. Contractor, and J. Hendler. The Web Science Observatory. *IEEE Intelligent Systems*, 28(2):100–104, Mar. 2013.

[31] M. Van Kleek, D. A. Smith, N. Shadbolt, et al. A decentralized architecture for consolidating personal information ecosystems: The webbox. 2012.

[32] M. G. Van Kleek, W. Styke, D. Karger, et al. Finders/keepers: a longitudinal study of people managing information scraps in a micro-note tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2907–2916. ACM, 2011.