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

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

Daniel Alexander Smith
Web and Internet Science
University of Southampton
Southampton, UK
ds@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

Keywords

Web Observatories, Personal Data Stores

1. INTRODUCTION

The concept of a Web Observatory [20, 10] 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 multi-purposable 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. [17]) 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 well-being [**?**].

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-centered design, to cultural anthropology, for example) but such repositories also represent unprecedented privacy risks and targets for identity thieves, misuse by commerical 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 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 first 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 promiting data sharing in such settings.

2. WHAT ARE WEB OBSERVATORIES

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

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 [9, 20]. 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 (PDSES)?

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 synchrnous 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. he 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 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. It is from this that PDS requirements have been derived by researchers working on such platforms (e.g. [**?**]). Such studies have documented the breadth and often idiosyncratic nature of people's personal information practices that have been 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 work around such limitations in order to manage their information archives.

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 well-being. 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 management (such as through sensed data streams representing the individual's vital statistics and activities).

With more sophsiticated functionality, such a personal web observatory might monitor one's social interactions, and correlate such interactions with states of well-being; 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 well-being, 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. TOWARDS DISTRIBUTED PERSON-CENTRIC WEB OBSERVATORIES

A next logical step from a PWO, then, is facilitating the interaction of multiple such observatories to enable analysis at larger scales and with the combined datasets of many individuals.

5.1 Early PDS concepts

However, this is a fairly minimal power which is hardly congruent with the increasing clamour concerning rights to data, including the spread of enforced transparency of data from the private sector [7] and the vogue for freedom of public sector information [16], and technology (and technology policy) together with new attitudes to transparency bring more possibilities. In the UK, a government initiative called *midata* [19] 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[2] and the French Mesinfos group[3]). 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.

Independent of such legislative approaches, both academic and industry-led efforts also began to commit resources to research towards identifying ways that end-user citizens might, in the face of the vast growing repositories of data being held about them, enjoy more control and privacy. An academic consortium known as *Vendor Relationship Management* (VRM) at Harvard's Berkman Center was realised to conduct multifaceted research into socio-legal-econo-technical approaches that might be employed. Among the products of this research was a vision that users might stand as their own information brokers, and start to act as peers with service providers, capable of negotiating fair and equitable mutual terms of data use during interactions with them[2]. Out of this work emerged the earliest mentions of Personal Data Stores for realising such capabilities in the context of online e-commerce, inspiring more than a dozen different Personal Data Store offerings, platforms and services backed by commercial start-ups since 2001 [1].

As an example, consider Mydex, whose proof-of-concept offering dates back to 2009 [12]. 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[4] — the ability to choose the location of the data centre in which the PDS is stored. Similar open source personal data storage containers include The Locker Project[5], data.fm[6], Owncloud[7], and OpenStack[8], each of which provides various degrees of easy-to-set-up 'personal cloud' software that can be used to store and host content on the user's own server on the Web.

A consistent theme of commentary in this area has seen Personal Data Stores (PDS) as important, if not essential, capability for end-users towards growing a healthier global "personal data ecosystem". For example, an independent study commissoned by The World Economic Forum documented ways that the value of personal data might be further "unlocked", citing Personal Data Stores as a core enabling mechanism to turn end-users from consumers into more autonomous data brokers[5]. A separate comprehensive analysis by *Ctrl-Shift* on emerging commercial PDS platforms and offerings projected an enormous economic opportunity for PDS services in the next five years[1]. 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.

5.2 Failure to Launch: Barriers to PDS Adoption

Yet despite the extensive needs analysis and market potential identified, early personal data store offerings have thus far failed to attract substantial attention from users. While a number of factors are likely responsible, so the lack of interest among users has been attributed to the fact that many of initial PDS platforms have sought to simply re-create existing end-user experiences offered by popular apps and Web platforms, rather than creating new functionality. Despite the benefit that these PDS offerings provide in terms of data security, users are often less compelled to try something new if the tangible experience nothing new, while data security remains an abstract, inestimable threat which does not necessarily easily compel behaviour change [3]. Finally, since the very purpose of PDS offerings is to protect user data from third party access, these platforms cannot derive revenue from user data and must resort to subscription models — always less attractive to new users than than offerings that are completely free to use.

[2] `www4.va.gov/bluebutton/`.
[3] `mesinfos.fing.org/`.
[4] `www.theguardian.com/world/the-nsa-files,www.ub.uio.no/fag/informatikk-matematikk/informatikk/faglig/bibliografier/no21984.html`.
[5] `lockerproject.org`.
[6] `data.fm`.
[7] `owncloud.org`.
[8] `www.openstack.org`.

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 [15], and it is important to challenge and resist it.

Heath et al write [12] that "there is market evidence that [the person-centric model of control over personal data] is starting to establish itself," but even they see a challenge to getting the model to work. Three conditions need to obtain simultaneously, on the account of Heath et al: PDSs must (i) make life simpler/better for the individual, (ii) appeal to data consumers by solving some of their problems (e.g. costs, or legal liability), and (iii) solve some pressing challenge that is holding back developers and entrepreneurs in this space. To these three, we can add a fourth, which is to rejig current data protection thinking. At the moment (2014), there are three key roles in the standard model of data protection: the data subject, the data controller and the data processor. The owner of a PDS is none of these (or none exclusively — he or she is likely to be all three at various times), and it is hard to see how individuals can exercise autonomous control over the data that affects them without some recognition of them as active agents in a different kind of role. Furthermore, data protection legislation is intended to cover cases of personal data being misused by others; it does not cover cases where individuals accidentally (or deliberately) identify themselves. Of course, this is a reasonable starting point for protection, but if it is the only principle, it means that if an individual 'takes charge' of his or her data, he or she *loses* the cover of Data Protection Acts.

6. SIX NOT SO EASY PIECES: CHALLENGES TOWARDS REALISING THE PDS VISION

The goal of providing individuals with the capacity to maintain their own information longitudinally imposes a number of challenges to supporting the kinds of information activities we have described. In particular, we see six broad categories of challenge to be met; the first, most fundamental of which pertains to effective *longitudinal keeping*. Enabling individuals to keep their data safely for a long time, while ensuring its continued accessibility and usefulness impacts both the data formats and methods used to store them. For example, since a person's physical computational hardware is likely to fail with age, methods need to be in place for ensuring robustness to such failures, such as multi-device replication and easy migration from older to new devices over time. Moreover, as evidenced by Moore's law [18], since the technical capabilities and properties of such data storage devices and platforms are likely to change fundamentally, PDSes must be designed to accommodate (and take advantage of) such changes as they arise. The devices and technologies that have made the PDS vision possible date back only a couple of decades, whereas a safe haven for data such as we are envisaging might well have to last a working lifetime (before we even consider the issues surrounding inheritance of data after a death).

A second challenge is allowing individuals who might have little or no experience in the intricacies of data management to cope with the burden of data security and longitudinal maintenance. Using current tools and services, for example, managing your data yourself still means taking pains to ensure that one's personal data is not lost to hardware and software failure, malicious attacks, or safely migrated to new platforms and devices; such efforts require vast investments of time, effort and expertise. A general lack of expertise or willingness to do this means that people currently rarely know how, or bother, to back up or consolidate their data. Thus it is no surprise that individuals have been motivated to outsource maintenance of their data to third parties, such as cloud providers. In order to facilitate autonomy from such services, therefore, PDSes must seek to support directly, and automate where possible, tedious data maintenance tasks that have plagued PC users for decades. Such automation could both ensure compliance for promoting data security and integrity, such as continuous backup regimes, thereby countering recent studies of the extremely low compliance of personal data backup and security maintenance practices [6, 8].

A separate set of challenges arises from the shift back from service-provider controlled data storage to a user-centered model of data management. Although this will re-empower users to control the organisation of their data spaces, and eliminate the pervasive problem of data fragmentation [13], [11], the challenge with the increased flexibility that this approach affords is that it requires re-consideration of how third-party applications and services can interact with such data, which have traditionally been pre-defined to operate on a fixed, typically application-provider established, set of data representation(s) and manipulations. In a consolidated, user-centric data model, on the other hand, such representations may be be specified or modified by the individual, or by some other third-party application(s) on behalf of them, and thus applications themselves must be designed to accommodate such variability among representations.

The need to comply with local, national and international data handling requirements pose a fourth set of challenges. In particular, if PDSes are to support the storage of identifiable information, or more critically, regulated sensitive information such as individuals' medical records, then PDSes must implement a variety of security standards (e.g. [**?**]) to ensure secured storage. Perhaps more difficult might be achieving compliance with the additional data handling requirements imposed by these regulations beyond how it is stored and encrypted; in particular, key handling requirements and guaranteeing aspects of physical access to the machine(s). The integrity of data must also be secured —

for instance, although a patient should have the right to challenge and correct inaccurate medical data, if the PDS is to store a version of medical data that is likely to be used (for example, in support of medical treatment in a foreign country), the data would need not only to be accurate, but also of appropriate provenance in order to be properly adapted to the standard workflows of medical treatment.

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

Perhaps the ultimate set of challenges, however, pertain to accommodating change as it affects both the information itself and the practices and activities surrounding it, over the years that a PDSes is intended to operate. Technologies that bring in new ways that data is used and generated seem to be introduced every quarter, placing new demands how this information needs to be accessed, created and used. The most recent examples include wearable computing and "always on" wearable sensor technology, from simple devices such as Fitbits [9] and Fuelbands[10] 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[11]. 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 [4] 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 [14].

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

[9]Fitbits - `www.fitbit.com`
[10]Nike+ Fuelband - `www.nike.com/fuelband`
[11]Google Glass - `www.google.com/glass`

Figure 1: A social healthcare application running on the the INDX Personal Data Store, showing data from multiple sources evaluated together.

7. DISTRIBUTED STORAGE OF WEB OBSERVATORIES

Additionally, there are operational challenges in implementing a distributed personal data store architecture. Specifically that of:

1. Distributed sharing
2. Distributed authentication
3. Distributed redundancy, synchronisation and replication

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.

The problems of authentication are that of trusted identification of users within a distributed system. In a traditional system, a user would login to a system in order to authenticate, however in a distributed system, the user would have to register for each individual remote system, and associate each of their accounts together. Instead, we can use distributed identification system such as OpenID in order to enable a single identity across systems.

The problems of synchronisation are that users who want to protect their data against catastrophic loss (i.e., from hardware failure), from network failure, or from unexpected intercept. A solution is to provide a system that automatically synchronises data between multiple redundant servers, and automatically fails-over to online servers when others go offline.

8. REFERENCES

[1] The new personal data landscape. Technical report, 2011.
[2] J. M. Agustin and W. M. Albritton. Vendor relationship management. 2001.

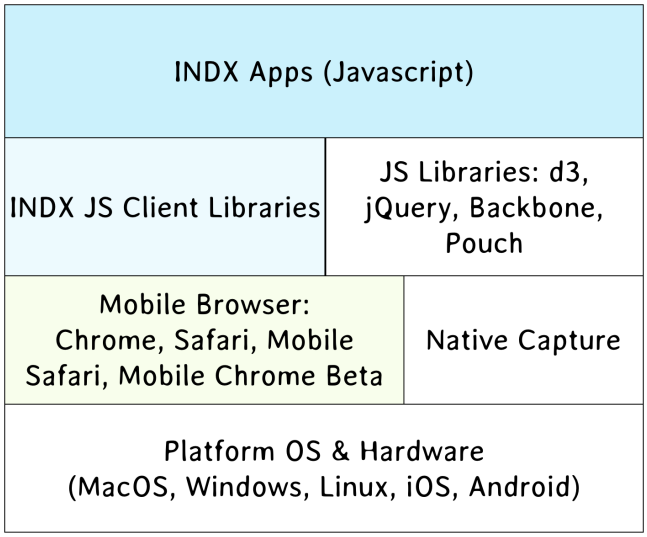

Figure 2: The INDX layer cake, showing how different technologies interact with each other.

[3] A. Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

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

[5] Boston Consulting Group. Unlocking the value of personal data: From collection to usage. Technical report, 2013.

[6] A. Chervenak, V. Vellanki, and Z. Kurmas. Protecting file systems: A survey of backup techniques. In *Proceedings Joint NASA and IEEE Mass Storage Conference*, volume 3, 1998.

[7] A. Fung, M. Graham, and D. Weil. *Full Disclosure: The Perils and Promise of Transparency*. Cambridge University Press, 2007.

[8] M. A. Grasso, M. J. Yen, and M. L. Mintz. Survey of handheld computing among medical students. *Computer methods and programs in biomedicine*, 82(3):196–202, 2006.

[9] 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.

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

[11] 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.

[12] W. Heath, D. Alexander, and P. Booth. *Digital enlightenment, Mydex, and restoring control over personal data to the individual*, pages 253–269. IOS Press, 2013.

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

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

[15] K. OâĂŹHara. Are we getting privacy the wrong way round? *IEEE Internet Computing*, 17(4):89–92, 2014.

[16] K. OâĂŹHara. The information spring. *IEEE Internet Computing*, 18(2), 2014.

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

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

[19] N. Shadbolt. *Midata: towards a personal information revolution*, pages 202–224. IOS Press, 2013.

[20] 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.