Becoming Dataware

Richard Mortier, Robert Houghton,
Anya Skatova, Christian Wagner,
Tom Lodge, Jianhua Shao, James Goulding
Horizon Digital Economy Research
University of Nottingham, UK
firstname.lastname@nottingham.ac.uk

Anil Madhavapeddy, Jon Crowcroft
Computer Laboratory,
University of Cambridge, UK
firstname.lastname@cl.cam.ac.uk

ABSTRACT
In Digital Economy circles it is increasingly common to speak of having a "personal contextual footprint", "owning data", and "exploiting the value in your data", as if these concepts were well defined and widely understood. However, how personal data should be treated, what specifically constitutes a "digital footprint", and how it can be valued and exploited are all open questions. Previous work in the Horizon *Personal Containers* project designed the Dataware framework within which "your data" can be seen more as a service than a good. This paper describes the recently initiated *Becoming Dataware* project that will explore these notions further, expanding techniques and technology, and exploring the behavioural psychology of ownership of personal data.

Keywords
Privacy, Personal Data, Identity, Ethics

1. INTRODUCTION
Understanding and exploiting the personal contextual footprint is one of the core themes within Horizon and the wider Digital Economy Programme more generally. The *Becoming Dataware* project examines the linkages, tensions and interplay between both technical infrastructure and human challenges within the theme. This also constitutes a vantage point from which to begin to consider more generally the potential for transformative societal impacts from 'personal data' developments we are seeing in the digital world.

Initially explored within the *Infrastructure pilot*, and subsequently developed through the *Personal Containers* project, the current status is that we have described, refined and implemented early prototypes of components in the Dataware framework (Fig. 1). Simultaneously, the *Homework* project (EP/F064276/1) has developed a novel home routing platform that enables more intuitive, more capable home network management through manipulation of network traffic flows (Fig. 2); and the *C-AWARE* project (EP/I000240/1) has developed infrastructure for collating, storing and representing household energy consumption data. Building on all of these infrastructure technologies, the present project has three main aims:

1. To investigate the novel features of the Dataware platform from a behavioural perspective;
2. Informed by this, to develop and extend the Dataware technology with the intent of a first deployment, focusing on two particular datasets: household and device network use and energy consumption; and
3. To use this framework and technology to explore more general features of the psychological aspects of "owning"/controlling access to personal and private data.

Clearly, to gain a rounded appreciation of these problems, a multi-disciplinary approach where study of human behaviour feeds into and is subsequently informed by development and deployment of new technology is required. The problems we address can be divided into three disciplines: Psychological, Infrastructural and Computational.

1.1 Psychological Challenges
One way of framing the psychological challenges in this area is to consider the decision making processes surrounding giving informed consent to the collection, sharing and processing of personal data; How do people actually conceive of and value their personal data and what does this imply in terms of utility and potential uses by the individual and others? By addressing these questions we can begin to understand which factors affect people's choices in relation to their personal data.

A subsidiary theme also pertains: how do people relate personal data their 'lived lives' and in particular their sense of identity and memory for what they have done? Over 60 years of psychological research has demonstrated that far from being veridical as is often assumed, memory is in fact volatile, unreliable and easily biased or influenced. Indeed, the efficacy of many health interventions (e.g., from food diaries to some aspects of psychotherapy) is based upon correcting these weaknesses. This suggests that the issue of personal data itself raises deep issues concerning not merely understanding its nature and form (in terms of files, uses, security and datamining outcomes) but also its potential content.

1.2 Infrastructural Challenges
To build on the results of these psychological explorations, and to ensure that technical developments can take advantage of and verify psychological understanding, we will develop the Dataware platform further. In particular, we will engineer the platform so that it can be deployed with users to enable real-world validation of understandings gained through techniques such as surveys.

This entails further engineering of the existing Dataware implementations, to build a deployable *Home Information Hub* (*HiHub*) that can collect personal data from real users. The specific real-world data types to be incorporated include home energy and network usage data, collected and managed via a Dataware-enabled Home Information Hub based on the Homework home router platform and the C-AWARE energy consumption data collection system.

1.3 Computational Intelligence Challenges
Making sense of your contextual digital footprints, and the value contained therein, is difficult. Tools to help users do so are required, supported by the infrastructure and designed in line with understanding of the many psychological factors around personal data. Such tools require mechanisms to map the value of potentially vague information (whether to numbers, intervals or

even distributions), and to handle the impact of potentially repetitive queries. This will enable creation of privacy metrics for data sources that are meaningful to users, and can subsequently be used to create a range of helper applications, able to e.g., alert users to, and protect them against, attempted privacy infringement.

2. OUTCOMES

We anticipate a range of useful concrete outputs from this project, in addition to the expected increase in understanding of how to manage people's personal contextual footprints.

The HiHub platform is being developed under open-source licenses, and so code will be available for others to use and build upon. Energy and network use datasets will be collected and made available, subject to appropriate ethical review.

Perhaps most interestingly from an inter-disciplinary research point-of-view, the proposed developments suggest the possibility of a new experimental paradigm enabled by the Dataware catalogue: *the ability to manipulate an experimental participant's data without the experimenter needing visibility into the data itself*. This creates the opportunity to understand both what can be inferred simply through the catalogue alone (e.g., the freshness or quantity of sources might be indicative of some features for some people); or, through "misdirection," to observe effects that might be more ethically difficult otherwise, such as determining probabilistically the point at which individuals notice missing records, or the interleaving of real records with false ones.

Experiments to be carried out within the 'memory' theme will determine probabilistically the point at which individuals notice missing records, or the interleaving of real records with false ones etc. In general we do not anticipate utilizing participant's personal data held prior to participation in our work. Further, initial studies will likely involve a phase where participants are asked to undertake a task or a behaviour that is explicitly about generating data for later manipulation so as to control the type and scope of information that might be collected. This information would then be presented for response from participants in a later phase of the experiment. In terms of the wider literature, the methodological space we intend to inhabit can be thought of as combining the work of Sparrow et al on digital offloading with some of the less controversial paradigms in false memory research [1].

This would in itself constitute a Dataware application, and so would represent a 'hard test' of the paradigm and the attendant ethical issues involved in allowing the production of such programs in future. These ethical issues will be explored through interaction with ethics committees across the University and, potentially, external experts. In terms of general theory, the discussion surrounding personal data and 'data-as-memory' has been afflicted with a tendency to 'folk psychology' notions of memory and self that are unsupported by the psychological literature. Thus we also hope work in this area will give an opportunity to discuss these issues to allow a transfer of knowledge from psychology into the personal data literature.

3. CURRENT WORK

Progress so far has focused on two fronts: psychological study into relative valuation of personal data privacy through online surveys, and technical development of the HiHub platform.

The first activity involves applying psychological scaling techniques to understand the relative value of various types of personal data (i.e. the implied value of retaining privacy) in different scenarios. The value of privacy is measured in terms of the money one would pay to avoid privacy invasion when using a service. Notionally, this mimics the choices already offered to users within the business models of various 'free' email services and similar: pay for privacy or use the service for 'free' as a donor of personal data for marketing use. Our experimental manipulation was thus to vary the price of retaining of privacy.

60 volunteers filled in an online survey via Qualtrics software. The survey constituted a randomized list of 84 items, each of them presenting a participant with a two alternative forced-choice (2AFC) question: either to pay for the service requested amount of money each month and secure their private information, or not to pay and share their private data. The amount of money they were requested to pay was manipulated from 50p to £20, and for each item we manipulated the type of private data they were requested to share. Examples of private data included bank statements, health records, household and mobile phone bills, Internet search history, and location data. Our intention was to see if participants could make principled, consistent and content-sensitive decisions; the null hypothesis was that the data as a whole would be chaotic and indeterminate, as some have previously suggested is likely.

The results demonstrated that majority of participants (>70%) are ready to pay as much as £20 a month to secure some pieces of information (e.g., bank statement and health records) and choose to pay nothing to secure others (e.g., online advertising click history and loyalty cards history). Some participants (~20%) are ready to pay to secure any type of private data, while others (~10%) do not want to pay anything to secure any pieces of their private data. This is suggestive of individual differences in terms of attitudes to sharing private data. In addition, the results show that people might not perceive private data in terms of their likely value to external parties, in line with a standard model of economic pricing. Rather, they make more categorical decisions based on other notions perhaps concerning personal sensitivities and anxieties. In the next study we aim to extend a similar design to other situations where people must share their personal data (e.g., applying for a credit card or visiting a website that promises a voucher or better services in return to private data), and investigate whether a similar pattern of results emerges.

Technical platform development is nearing the end of its first phase. We have selected the deployment platform (the Marvell *DreamPlug*, a small-form-factor plug ARM-based PC), and ported code released through the Homework,[1] Dataware,[2] and C-AWARE[3] projects. We have integrated and extended these platform elements to create a wireless access point that can record network and electricity use, and expose those datasets as Dataware providers. We are now beginning to plan initial deployments through which we will gather data enabling the computational intelligence strand to begin, and the psychology strand to move to its second phase, using real data rather than simple survey techniques.

ACKNOWLEDGMENTS

This work was funded by the RCUK Horizon Digital Economy Research Hub grant, EP/G065802/1.

[1] http://github.com/homework
[2] https://github.com/horizon-institute/dataware.*
[3] https://github.com/caware

REFERENCES

[1] Sparrow, B, Liu, J., & Wenger, D.M. (2011). Science, 333, 776-778.

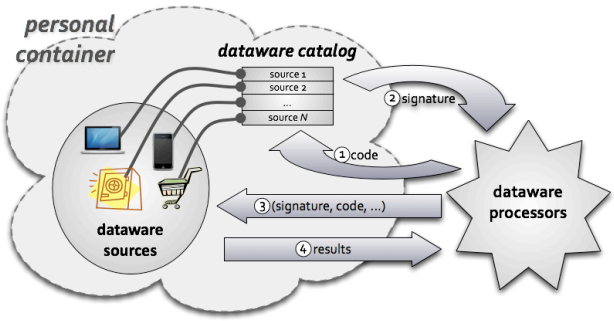

Figure 1. The Dataware architecture. The *catalog*, under the user's control, contains information indicating the user's various personal data *sources*. *Processors* attempt to initiate processing of users' personal data by requesting credentials from the catalog. If the catalog permits access, it grants credentials to the processor enabling it to install the processing it wishes carried out.

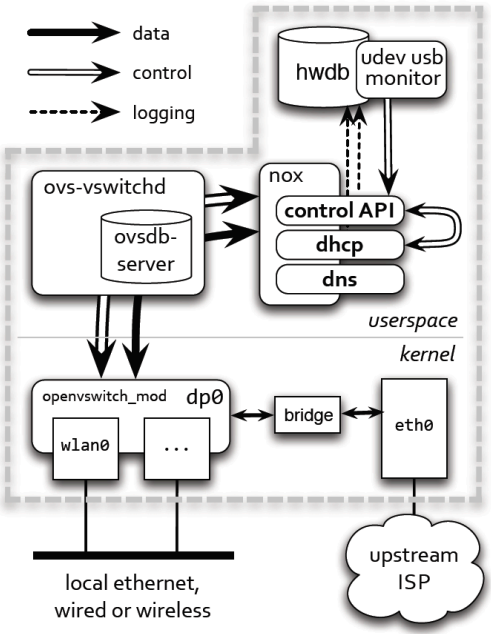

Figure 2. The Homework Router architecture. Open vSwitch and NOX manage the network interfaces. Three NOX modules provide a web services control API, a custom DHCP server, and a DNS proxy. The HiHub builds on this, adding extra network and energy consumption logging facilities, and incorporating Dataware Provider functionality.