DEMO: Dataware, A Personal Data Architecture

Richard Mortier, Tom Lodge,
Neelam Bhandari, Robert Houghton,
Anya Skatova, James Goulding,
Christian Wagner
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

Dataware is a set of technologies developed via two Horizon hub funded projects, *Personal Containers* and *Becoming Dataware*. Part of Horizon's mission to address the challenges posed by the lifelong contextual digital footprint, Dataware enables individuals to regain control over use of the digital data constantly being created about and by us. The essential enabling design feature is that Dataware provides a set of mechanisms that enable you to control access by others to your data held in multiple locations according to how it is generated, e.g., at banks, retailers, utility companies, broadband providers or online social networks.

Having previously described the Dataware architecture [1], we have been working to refine its design and implement its various components, along with some exemplar applications. We will demonstrate a prototype **home information hub** which collates data from sources in the home (energy consumption monitoring and home network usage), and makes it available for processing by clients using Dataware.

1. THE HOME INFORMATION HUB

We envision the Home Information Hub (HIHub) (Figure 1) as a central point within the home that acts to collate, manage and store data generated by both householders and devices about the home. As a first prototype, we have configured a small form factor PC (a Dreamplug) as a wireless router, running *Open vSwitch* (http://openvswitch.org) to provide a programmatic extension to the router's network stack, and *NOX* (http://noxrepo.org), a component that interacts with Open vSwitch and allows us to hook into the data flowing through the router.

Using NOX we capture DNS – the protocol invoked to translate Internet names to addresses in, e.g., URLs – allowing us to monitor the sites and services that devices connected to the HIHub. The HIHub also acts as a sink device for the Current Cost energy monitor, allowing it to collect electricity consumption of the home and any devices plugged in via Current Cost device monitors. Both data sources (network service use and energy consumption) are provided as independent Dataware resources via software running on the HIHub and in the cloud. Finally, we provide several other data sources as Dataware resources, including browsed web-page content and social network data (Facebook, Twitter).

Figure 1: Prototype Home Information Hub.

2. DEMO SCENARIOS

The demo consists of a running HIHub collating data and presenting it as Dataware resources. We then present two forms of Dataware client. The first is a simple set of web pages providing an audit interface to view both your data available as a Dataware resource, and accesses made to it. The second is representative of a practical application of Dataware that we anticipate will be relevant to the audience of DE 2013: The Total Recall Experiment.

Although your personal contextual digital footprint is viewed in many ways as at the heart of the Digital Economy, and is a key feature of the ubiquitous computing revolution, it can be difficult to carry out research into people's perceptions and behaviours around their digital footprints. Ethical considerations coupled with the present need for researchers in this field to engage very directly with a wide range of personal data for each subject make gaining clearance for such experiments problematic. To demonstrate the benefits of the Dataware approach over traditional methods, we have developed an example of a simple recall experiment that uses Dataware.

At its core, the experiment consists of setting subjects an image search task which they carry out using a browser that records the URLs they visit. The subject is then presented with a sequence of single URLs and asked to indicate whether or not they in fact visited each URL. Most of the URLs they are presented with are ones they visited, but some URLs that they did not visit are randomly placed in the sequence. In this context, the purpose of such an experiment is to begin to understand how well subjects can recognise their personal digital footprint – in this case, their web browsing history.

Traditionally carrying out this experiment would require the experimenter to setup a customised browser and run people through in lab conditions; to carry it out under more natural conditions would require the experimenter to have unfettered access to each

DE '13 Salford, Manchester, UK

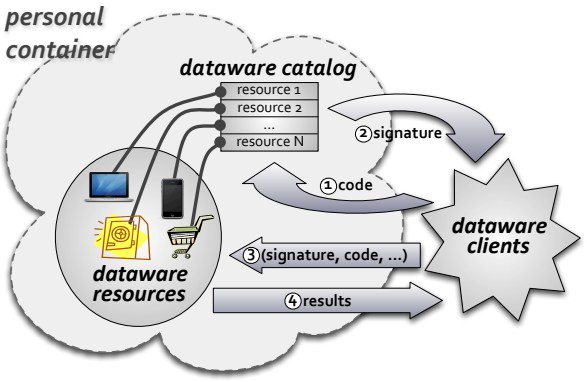

Figure 2: Dataware personal data architecture.

individual's browsing history. The overheads in both cases are high, and in the second case, the level of access to personal data required may be ethically problematic. Using Dataware, this experiment can be carried out repeatedly, at scale, under more natural conditions, and without the experimenter requiring direct access to the browsing history of the subject:

1. The experimenter sets up as a Dataware client, which simply means that they run a web site supporting the necessary Dataware protocol URL endpoints, and then encode their experiment as a piece of data processing code within the Dataware framework. In this case the code would simply retrieve and present (say) the last day's edit browsing history with some URLs removed and some added.[1]
2. Hearing about the experiment, a subject[2] chooses to visits the experimenter's website, and clicks a link indicating their desire to participate. This causes the experimenter's web site to contact the subject's Dataware catalog requesting permission to run the experiment against the subject's browsing history.
3. The subject is notified that there is a pending processing request, by being redirected by the experimenter's site to their catalog in this case.[3]
4. Assuming they are willing to take part, the subject consents, which causes two actions. First, the experimenter's site receives a token which it can present to the subject's browser history resource along with the code representing their experiment. Second, the subject is redirected back to the experimenter's site to be presented with the processed list of URLs representing their browsing history over the specified time period.
5. The subject then completes the experiment by indicating whether or not they visited each displayed URL in turn. The experimenter records only whether the correctly recalls whether or not they visited each URL – the specific URLs are not recorded.

[1]Clearly some processing of the complete URL history would be appropriate as users generally do not remember all the details and parameters of each URL, even if they do remember the site or service accessed.

[2]Assumed to have already installed required Dataware components, in this case a catalog and a registered browsing history resource.

[3]In scenarios where the subject is not immediately present at the point the request is made, deferred processing of the request may be required with notification occurring out-of-band by email, phone or similar mechanism.

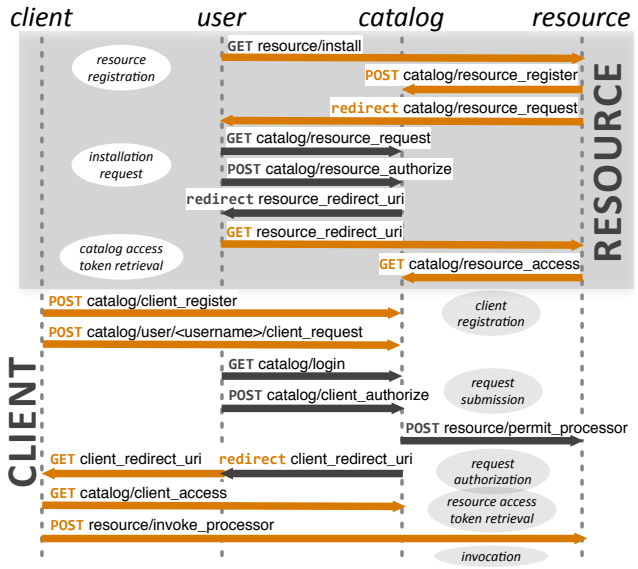

Figure 3: Dataware Protocols for a resource to register with a catalog, and for a client to access a resource.[4] Based on OAuth 2.0 Authorization protocol v22. Interactions the developer must handle in the app are in orange; those in grey are either automatic or part of normal web browser behaviour.

3. TECHNOLOGIES USED

Figure 2 depicts the Dataware architecture in more detail: your **catalog** represents your wishes in this system, implementing any static policies you may have over access to your data, and providing a channel by which to communicate directly with you in cases where static policy is inadequate; your **resources** maintain fragments of your data, typically mapping to individual data sources; and **clients** represent other parties that wish to compute over your data. Resources and Clients implement the Dataware Resource and Client Protocols respectively, depicted in Figure 3.

4. REQUIREMENTS

We will provide the necessary hardware (an energy monitor inductive clamp and display; a dreamplug small form-factor PC; and a laptop for display). We will require table space (say 1x2 m^2), poster hanging space, and access to mains power (4 sockets) and the public Internet, unproxied if possible (to access the cloud components of the service). Access to a larger (e.g., 32") display screen is desirable.

Acknowledgements. This work is supported by Horizon Digital Economy Research, RCUK grant EP/G065802/1. Packages and source are available under open source licenses at http://github.com/horizon-institute – search for repositories beginning `dataware.`.

5. REFERENCES

[1] D. McAuley, R. Mortier, and J. Goulding. The Dataware Manifesto. In *Proc. 3rd IEEE International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–6, Bangalore, India, Jan. 4–8 2011. Invited paper.

[4]Documented at https://github.com/horizon-institute/dataware.catalog/wiki/.