Leveraging the Power of the Cloud to Create a Marketplace for Crowd-Sourcing of Personal Data

Dominic Price
Horizon Digital Economy Research,
University of Nottingham,
Nottingham, NG7 2TU, UK
Email: firstname.lastname@nottingham.ac.uk

Bio

Dominic Price is a Computer Science Research Fellow at the Horizon Digital Economy Research Hub (http://www.horizon.ac.uk) at the University of Nottingham; he has been a member of Horizon since 2010 and a member of the University of Nottingham since 2007. His current research interest is in crowd-sourcing, in particular computer support of crowd-sourcing and software toolkits that aid human-computer interaction in crowd-sourcing. Previously he worked in the Mixed Reality Research group at the University of Nottingham and has worked with many large industrial partners including Microsoft, the BBC and BT as well as smaller artistic partners including the artist groups Blast Theory and Active Ingredient. Prior to his research career he worked in industry as a software developer and technical manager for a small eCommerce firm after completing his undergraduate degree in Computer Science in 2004.

Abstract

In this paper we briefly describe what is meant by the term ‘crowd-sourcing’ and present our vision for a software toolkit that utilises cloud technology to implement a platform and marketplace for crowd-sourcing, in particular the crowd-sourcing of personal data such as GPS trails and health information. We discuss the software architecture that we are implementing as a prototype of such a toolkit and give an example of a proposed crowd-sourcing activity that will use this toolkit to show how the toolkit may be used by third parties.

Introduction

Crowd-sourcing is a relatively new term that encompasses an extremely large variety of tasks that aim to use human intelligence and activity on a large scale to gather data or to solve predefined problems [1]. Although there have been efforts to create a more complete definition of crowd-sourcing [2] it is generally used as an umbrella term for any number of different tasks. In this paper we will use the term ‘crowd-sourcing’ in its broadest sense, that is any task that recruits participants with an ‘**open call**’. An open call is a request for participation that requires potential participants to respond to the call as opposed to participants being specifically recruited. As the call is open there could potentially be an unlimited number of responders with any background though calls can be targeted to specific groups and closed once a sufficient number of participants have responded.

The following are some examples of specific types of crowd-sourcing activities/applications to illustrate the breadth of crowd-sourcing:

- Human intelligence task – An activity in which a human performs a task in which automated computation typically cannot provide as high a level as accuracy as a human. Image classification is one example of this, the Galaxy Zoo [3] project uses

crowd-sourcing to classify images of galaxies taken by the Hubble space telescope, a task in which automated methods produce less reliable results than humans performing the same task.

- Citizen Science – Members of the public are used to collect and analyse scientific data on a scale larger than that of small teams of scientists would be able to manage alone (note, Galaxy Zoo may also be categorized as citizen science). The Royal Society for the Protection of Birds (RSPB) [4] operate a yearly **Bird Watch** in which members of the public are asked to make records of bird sightings and to submit the results to the RSPB. The RSPB are then able to analyse the results to look at changes in bird populations between years.
- Incidental crowd-sourcing – We define incidental crowd-sourcing as an activity that produces some useful outcome as a by-product or side effect of a person performing that activity for another purpose. reCAPTCHA [5] is one of the most famous examples of this type of crowd-sourcing. As a by-product of websites that use reCAPTCHA attempting to ensure that form submissions are being made by a human rather than an automated means, words from digitized books that have not been correctly recognized by optical character recognition techniques are digitized by the human submitting the form.

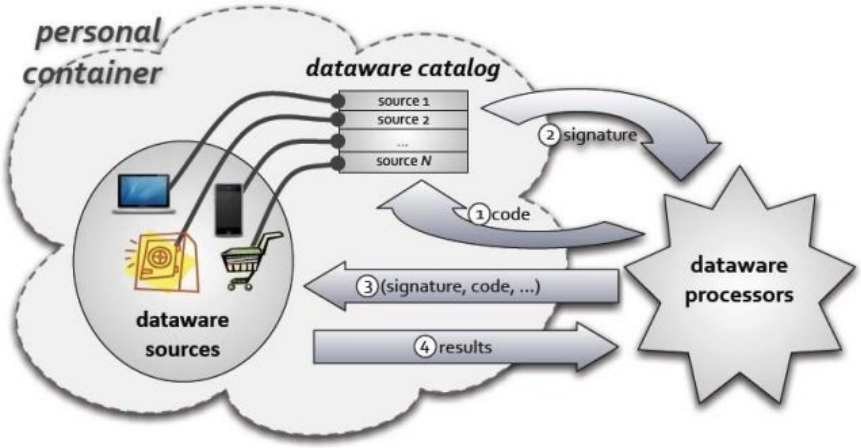
As can be seen from the above examples, crowd-sourcing is very broad and the categorisations between different forms of crowd-sourcing can be very blurred. As such creating a universal platform that could support all of the activity types would involve a huge software engineering effort. There do exist platforms for crowd-sourcing but these are typically more limited in scope, some examples are Amazon's Mechanical Turk [6] which provides the functionality for human-intelligence tasks and the Ushahidi [7] platform which is predominately used for citizen journalism style crowd-sourcing activities. There are however commonalities that cross-cut through the majority of crowd-sourcing activity types, these are:

- Recruitment – the process by which a call is sent out and participants are registered in the crowd-sourcing application.
- Incentivisation – the method by which participants are incentivised to participate, examples are competition, kudos and remuneration.
- Data storage and access – the physical storage for data gathered during a crowd-sourcing activity and the technical methods available to access that data.

It is these three aspects that we have taken in designing our infrastructure for a crowd-sourcing toolkit which is described below.

Personal Data and the Dataware Manifesto

Briefly, before we describe our proposed toolkit architecture we will introduce the concept of personal data and the Dataware manifesto [8]. The Dataware architecture (shown in figure 1) is the output from another of Horizon's research projects looking into how personal data may be stored and accessed in privacy preserving ways. Personal data is contained in any number of Dataware sources or, 'stores', for example health data would be stored in a 'health-store', a user may manage their own store or delegate this to a trusted third party provider. A **catalog** maintains a list of a user's stores and acts as a trusted broker to allow access to the data by third parties known as **Dataware processors**. Dataware processors send signed queries via the catalog to an individual's stores that the user can then opt to allow access or not before the query results are sent back to the processor in an anonymised form. The Dataware architecture underpins our proposed crowd-sourcing architecture.



Figure

Architecture

Our vision of a universal toolkit for crowd-sourcing of personal data is not one of supporting every kind of crowd-sourcing interaction 'out-of-the-box', as stated above this is a daunting task. Instead we visualise a toolkit that supports the basic activities that cross-cut crowd-sourcing (recruitment etc) and allows new crowd-sourcing activities to be developed in a modular fashion. We envisage that these modules could be shared and traded, perhaps for monetary value, thus creating a marketplace for crowd-sourcing. The visualised toolkit would then allow there to be a greater degree of separation between those groups who wish to crowd-source (the **Initiators**) and those groups that wish to develop crowd-sourcing activities (the **Developers**). Initiators will be able to assemble crowd-sourcing workflows from existing modules with little to know software development knowledge, whilst developers will be able to trade the modules that they develop.

Figure 2 below shows a high-level view of the proposed architecture and the major interactions between the different services and actors. The 'heart' of the toolkit is the Crowd-Sourcing Factory {1}; this is a Windows Azure cloud application that provides the front-end for the Initiators {5} and the Developers {6}. Developers can log in and create modules {2}; modules are small applications that perform a crowd-sourcing activity, for example creating a survey for participants to fill in. Modules are developed in a scripting language (currently only Python is supported) that works with an application programming interface (API) provided by the crowd-sourcing factory. Developers can share, trade or sell these modules with other developers and with initiators.

Initiators log into the crowd-sourcing factory where they can assemble a crowd-sourcing workflow from the available modules. Once they have performed this task, they can activate a crowd-sourcing *instance* {3}. The crowd-sourcing factory creates a new Azure instance in

the cloud which provides a harness to run the particular crowd-sourcing application instance along with storage for data collection. Thus, this instance is conceptually separated from any other instance and the factory providing a more secure environment than running the instances in one Azure instance. It also allows better auditing for individual instance usage which could, for example, be used to provide more accurate billing based on usage.
Once the crowd-sourcing instance is live, the next phase is recruitment; that is, the process of driving participants to the front-end of the instance so that they can participate. There are many ways in which this could be achieved and we envisage having a pluggable mechanism that is similar to the task modules that allow different recruitment paths to be used. For instance plugins could be developed for recruiting through social media such as Facebook or through more traditional means such as email.

Figure
Participants {7} who respond to the call can register on the instance site with an OpenID [9], no personal information is kept by the instance unless the participant specifically allows the release of such information through their catalog {4}. As a participant can have many different OpenIDs this can make it easier for a participant to maintain their anonymity across many crowd-sourcing instances. The participant also needs to provide a catalog address for access to their private data. The catalog will report what stores the participant has available, if suitable stores are not found the participant will be prompted to set up the needed stores otherwise they will not be able to participate. Once this process is complete the participant may then actually follow the crowd-sourcing workflow before the stores push data back to the crowd-sourcing instance. One the crowd-sourcing phase is complete, the initiator may then request the data from the instance and then close the instance down.

Current Status

We are currently in the final stages of development of the crowd-sourcing toolkit and the Datasphere integration with the aim for a public release at the end of April 2012. Once development is complete we will be running a trial that aims to crowd-source geographic and mobile device signal quality on train journeys.

Acknowledgement

The research on which this paper is based was funded by the RCUK supported Horizon Hub, EP/G065802/1.

References

[1] Jeff Howe (June 2006). “The Rise of Crowdsourcing”, *Wired Magazine.* Available online at http://www.wired.com/wired/archive/14.06/crowds.html [Accessed 22 August 2011].
[2] Estellés-Arolas, E. and González-Ladrón-de-Guevara, F. 2012. Towards an integrated crowdsourcing definition. *In Journal of Information Science XX (X) pp. 1-14.*
[3] Galaxy Zoo. http://www.galaxyzoo.org/
[4] RSPB Bird Watch. http://www.rspb.org.uk/birdwatch/
[5] reCAPTCHA. http://www.google.com/recaptcha
[6] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome
[7] Ushahidi. http://ushahidi.com/
[8] McCauley, D. Mortier, R. and Goulding, J. 2010. The Dataware Manifesto. *In Proceedings of Third International Conference on Communication Systems and Networks.* 4-8 January, Bangalore, India. DOI=http://dx.doi.org/10.1109/COMSNETS.2011.5716491.
[9] OpenID. http://openid.net/