A Framework for Crowd-Sourcing Personal Data

Dominic Price, Chris Greenhalgh, Shakir Ali, James Goulding and Richard Mortier
Horizon Digital Economy Research
Nottingham Geospatial Building
University of Nottingham

Nottingham NG7 2TU UK
+44 (0) 115 846 8923

{firstname.lastname}@nottingham.ac.uk

ABSTRACT

In this paper, we present a proposed architecture for a software toolkit for crowd-sourcing personal data from crowd-sourcing participants. We describe a priming study on the use of crowd-sourcing English language in context and how this has been used as motivation for the crowd-sourcing toolkit. Finally we consider the potential uses of such a toolkit.

Categories and Subject Descriptors

D.0 [**General**]

General Terms

Design, Experimentation

Keywords

Crowd-sourcing, personal data, cloud computing

1. INTRODUCTION

Crowd-Sourcing is the term used to cover a broad range of tasks that use human intelligence on a large scale to solve a predefined task. The important aspect that defines a task as a crowd-sourcing task is that it be open to an undefined group of participants [1]. Some examples of crowd-sourcing tasks are: human-computation – using the crowd to perform a computation task that is too complex for a software algorithm to compute, e.g. classifying telescope images of galaxies; citizen science – using volunteers to study, measure and observe on a mass scale, e.g. recording sightings of rare birds.

2. ENGLISH LANGUAGE IN CONTEXT PRIMING STUDY

The English language in context study is an ongoing project looking at the use of crowd-sourcing techniques to help develop improved contextual dictionaries as an aid for foreign language speakers. This study was developed as collaboration between Horizon Digital Economy Research (Horizon DER) and the Centre for Research in Applied Linguistics (CRAL) at the University of Nottingham. There is a dual purpose of the priming study, from the CRAL perspective it is to determine whether crowd-sourcing is an effective technique for gathering the information needed to produce a better contextual dictionary; and from the Horizon DER perspective it is to motivate and inform the development of a crowd-sourcing.

2.1 The Study

To date, there have been two phases in the pilot study. The first (Phase 1A) involved recruiting a group of 5 non-native English speaking volunteers to watch a 20 minute video clip of an academic lecture. The volunteers were then presented with a list of 30 phrases and asked they heard during the video. The purpose of this exercise was to test whether crowd-sourcing key contextual phrases is an effective way of collecting information. The preliminary results were compared against phrases that were known to be in the lecture and the initial results suggested that this could be a valid method of data collection. Survey information was gathered through a simple web application running in a web browser on an Android smart-phone and hosted in the Cloud on Microsoft's Windows Azure platform.

The second phase (1B) is an extension of phase 1A. For this trial 40 volunteers were recruited from a mix of native and non-native English speakers who were shown the same video as in the previous trial and again asked to select from a list of phrases that they believe that they heard in the lecture. In addition to this, the participants were also given the option to submit phrases not listed that they believed were important phrases they had heard in the lecture. Again an Android smart-phone was used to collect the data, this time using a native Android application to record and transmit the data to the web server for storage. This application also recorded whether phrases had been selected and unselected before submission as a metric of 'hesitancy'. The results of this trial are still pending analysis.

3. CROWD-SOURCING PERSONAL DATA

Crowd-sourcing often involves the participants revealing some personal information about themselves. For example in the case of the English Language in Context trial, participants need to identify their primary language; also, when the trial reaches the next phase and is run on a larger scale in actual lectures, information about participants attendance in lectures can be inferred from their responses. Another example of personal information is the tracking of location: within Horizon DER there are a number of projects that wish to track individuals' movements, but this is information that participants may wish to keep private as it could be used to identify a single person, their habits and even their relationships with other people.

A partner project to the Crowd-Sourcing Toolkit project within Horizon DER, the Personal Containers project [2], is developing

Digital Engagement'11, November 15–17, 2011, Newcastle, UK.

an application (the Datasphere) that will enable users to track and manage access to their personal information. Conceptually, this can be envisaged as a database that holds all of a user's personal information (though this information may in reality be stored in many disparate ways) and that allows third parties to ask questions about that information. The types of questions that are allowed to be asked will be vetted and managed by the system and will provide mechanisms for preserving privacy and for a user to maintain ownership of their personal data at all times. Further detail as to the mechanisms employed by the Datasphere to implement these features are outside the scope of this paper.

3.1 Crowd-Sourcing Toolkit

As crowd-sourcing is a term that covers more than one single task, and there are many different workflows that a crowd-sourcing task can consist of, developing a single universal application that supports all types of crowd-sourcing is a daunting prospect. Instead, we propose that there are a number of generic elements of crowd-sourcing that can be drawn out and an infrastructure developed to support these. In particular, participant recruitment, registration and management are common to most if not all crowd-sourcing activities. Our proposed toolkit contains support for all these interactions.

As stated above, creating a universal application to support all crowd-sourcing activities is a huge engineering task. What we instead propose is a 'marketplace' for crowd-sourcing activities (in essence, crowd-sourcing of crowd-sourcing). Our toolkit will provide an infrastructure in which crowd-sourcing modules that support a particular activity can be developed and then different crowd-sourcing workflows constructed by combining different modules. These modules can then be shared with other users to facilitate their crowd-sourcing activities. For example, in our English Language in Context trial, the activity being performed can be classified as a survey, and our toolkit can be used to develop a 'survey' module for use by that trial as well as other activities such as crowd-sourcing information on people's shopping habits. This enables non-programmers to reuse existing modules in creating new crowd-sourcing applications.

3.2 Dataware Architecture

At the heart of the toolkit is the crowd-sourcing factory (see figure 2). This is an application running in the cloud, providing the front-end interface for crowd-sourcing administrators (the user group which requests crowd-sourcing activities) and developers (the user group which develops crowd-sourcing modules) to log in and create crowd-sourcing modules and create an activity from selected modules. Once the workflow has been defined, the factory generates a crowd-sourcing *instance*, a separate application that is sandboxed from the factory and all other crowd-sourcing instances. This instance contains all of the necessary functionality for recruiting and managing crowd-sourcing participants as well as some storage for storing the results of the activity to be retrieved by the administrator once the activity ends.

Personal data is sourced from a participant's Datasphere: the crowd-sourcing instance knows the data it must request and talks to the participant's *catalog* to determine where that information is kept. The catalog authenticates the request and notifies the datastore that the crowd-sourcing application is allowed to access it.

A datastore is a repository for storing and managing access to personal information, each user has their own logical instance so that their data is firewalled off from other users. In the Datasphere manifesto, different types of information will be contained within different datastores, for example social information from social networking sites will be in a separate store to information about a user's energy usage. Each datastore is responsible for publishing a query interface. The instance then asks questions of the datastore in an appropriate format.

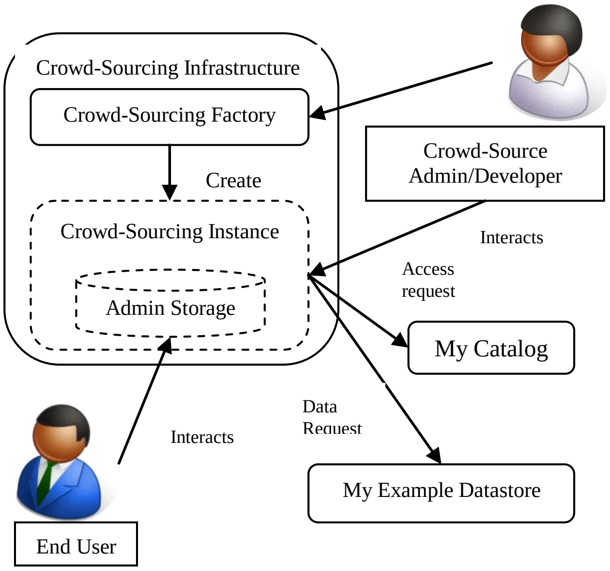

Figure 2. Toolkit Architecture Overview

In the example of the English trial, the catalog would direct the crowd-sourcing application to a *survey store*. The survey store would have either been previously set up by the user or the catalog would prompt the user to set up a new store from known providers (if no survey store exists the user will not be able to participate in the activity). The crowd-sourcing instance could then ask the required questions (what phrases were heard) and the survey store would in turn ask the participant for their results, once complete the store would return the results to the instance (potentially in a privacy-preserving obfuscated manner). The user will also be asked by the store to explicitly authorize the request in order to comply with regulatory requirements and the need to preserve users' privacy. This is the mechanism that allows personal information to be crowd-sourced in a way that maintains a users control over their own personal information.

4. ACKNOWLEDGMENTS

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

5. 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] 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.