KNOWLEDGE MEDIA
KMi
INSTITUTE

On the Privacy Implications of Releasing Consumer-Activity Data

Technical Report kmi-13-02
September 2013

Keerthi Thomas and Mathieu d'Aquin

The Open University

On the Privacy Implications of Releasing Consumer-Activity Data

Keerthi Thomas
The Open University
Walton Hall
Milton Keynes, UK
keerthi.thomas@open.ac.uk

Mathieu d'Aquin
The Open University
Walton Hall
Milton Keynes, UK
mathieu.daquin@open.ac.uk

ABSTRACT

There is growing awareness among web users that online organisations are collecting vast amounts of information about them and their activities. With this awareness is an implicit expectation that such data, generally called consumer data, should also be made accessible to the users themselves, and for their own benefit. Generally, it is considered not only fair that such data is not kept locked and out of user's reach, but also that it would lead to greater transparency and accountability of the organisations collecting them. As with any process which publishes data, there is a strong expectation that eventually it would lead to potentially complex privacy issues. In this paper, we focus on what we believe is significant and yet the least explored data type: consumer-activity data, i.e., data (Web access logs) generated by an organisation which tracks the usage and interactions of its online services and resources. We conducted a qualitative study with the goal to investigate the consequences of making such consumer-activity data available to the users who generated them, especially on the privacy challenges that might emerge, both from the organisation's point of the view and that of the individuals who were being tracked. This was achieved by exposing a group of 12 users to a prototype personal analytics dashboard, giving them access to information about their own usage and interactions with the online systems of a large educational organisation (The Open University in the UK). Although, the findings of the study showed that there are potential benefits for the users, it also identified several privacy risks and challenges which needs to be addressed.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

KMi-13-02 Technical Report, The Open University, UK.
.

General Terms

Privacy, Emperical Study

Keywords

Consumer-activity data, privacy

1. INTRODUCTION

Organisations, both the ones which model their business around free web based services (e.g. Google's search engine, Facebook's social networking service, etc.) as well as more traditional institutions with a significant web presence, indirectly collect vast amounts of data about their users and exploit them in a variety of ways, either to improve the user experience and indirectly benefit or to directly benefit their own business (advertisements, collective filtering, etc.)

From a user's perspective, such self-serving exploitation of user's data is not always seen in a positive light, thus new initiatives have emerged and are pushing towards greater transparency and openness, beyond the basic data protection practices. In the UK for example, the government is leading a consumer empowerment strategy in collaboration with leading businesses and consumer groups to give individuals more access to, and control over the data, companies hold about them [3]. While commercial sector players such as supermarkets, mobile network operators, banks and other industries are being persuaded to publish their consumer data, some of the giants of online services such as Google and Facebook are already leading by example. Google in particular is implementing its data liberation principle [2], to allow its users to download the data they create using Google's online services. Similarly, Facebook allows its users to download an archive of their online interactions as an 'activity log' [1].

With such top-down approaches to releasing consumer data evolving quickly and becoming more prominent, there are important questions from a user's perspective which remain unanswered: *Are these data really beneficial to the users and what do they think of it? If there are user-specific benefits, what is it likely to be (apart from sample use-cases)? By publishing these datasets, do both users and organisations face any risk to their privacy and security, and what are they? What is its impact on privacy policies?*

In this paper, our goal is to provide answers to some of these questions, focusing in particular on one of the most significant, and yet most under-explored part of consumer data: *Consumer-activity data*, which here refers to all data produced (Web access logs) as a side effect of a user's interaction with the websites or online software systems of a given organisation. To achieve our objectives we had to experiment by exposing real users from a large organisation with a complex web presence, to the potential mechanisms through which they could not only access but also visualise their own consumer-activity data. To this end, we designed a user study at The Open University (OU) relying on data collected from the logs of the institution's multiple online systems which captured the online activities of a group of 12 users (researchers, students, tutors, admin staff), for about 1 month, and provided them with access to such data through a dedicated 'personal analytics' tool.

We report here on our findings, namely, the benefits in releasing the CA data, privacy risks that are applicable to both users and organisations and the impact consumer-activity data has on the privacy policies which govern its use. We also discuss the implications of our findings beyond the specific environment in which the study is realised.

2. BACKGROUND AND RELATED WORK

In educational institutions, many systems store data about the actions of students, teachers and researchers. In this regard, JISC [22] mentions two types of data - (i) user activity data - a record of a user's actions on a website or software system or other relevant institutional service, and (ii) attention data - the record of what a user has viewed on a website or software system or other relevant institutional service. Both these data are similar to consumer-activity data, in that they are created from the user's interaction with an online system or resource. However, the difference is that consumer-activity data encompasses both user's 'action' and 'attention' on an online system. In addition to this, unlike JISC whose focus is on educational institution, consumer-activity data universally refers to 'all' types of organisations – academic, commercial and those belonging to the government.

In the UK, software systems which collect data are obliged to provide protection under the Data Protection Act if the data they collect relate to and contain information about an individual's identifiable attributes (e.g. name, date of birth, etc.). ICO [20] highlights such data as being of two types – *personal data* and *sensitive personal data*. Personal data relates to a living individual who can be identified from those data and personal-sensitive data refers to personal data pertaining to racial or ethnic origin, political opinions, religious beliefs, memberships of organisations (e.g. trade union), physical or mental health or condition, sexual life, convictions, etc. By nature, consumer-activity data are associated with an individual and therefore contain information that identify with the user who created it. The *midata* project [3] refers to such consumer-activity data as being personal data. Since the data profile from each software system can be different depending on the type of functionality they support, it is difficult to explicitly state beforehand what type of privacy issues one might encounter and the type of privacy protection users might need.

Apart from the Data Protection Act, the current trend of publishing consumer-activity data is driven by at least two principles stated in the OECD guidelines [34], one of them being the *Openness* principle which states that organisations should have a policy of openness about "developments, practices and policies with respect to personal data". In other words organisations are expected to be transparent in the way they collect and process personal data. Another OECD's *Individual Participation* principle states that individuals have a right to have visibility and access to their data and if required they should be allowed to rectify, correct, complete or erase the data held by others. This also ties in with the *Integrity/Security* and *Access/Participation* principles of FIP [16] which mandates organisations to make personal data visible to individuals so that its accuracy and completeness can be contested if necessary.

As a prolific collector of consumer-activity data, Google has been exploiting the data to not only power its Web analytics service [17] but also its targeted advertising. While some have evaluated the role of Google Analytics in improving the usability of e-commerce websites [19] and library services on the Web [42], others have shown how these analytics can be extended to measure and improve the performance of websites [37, 38]. Other parallel research efforts such as [14] and [15] concentrated on independently collecting and visualising web usage as user-centric analytics purely from a technical and architectural point of view, with one exception [12] where privacy and trust is briefly analysed. Focusing on the distributed leakage of personal data from user's interactions across a wide variety of websites, some have demonstrated how "privacy footprints" can be measured and analysed [25].

Qualitative research such as [27] captured Facebook activities to analyse its mobile privacy implications, in another similar study [26], data from location-tracking activities were used to elicit privacy concerns, in both works the emphasis was on the effect of user's mobility. Others works such as [7] have investigated how detailed tracking of user interaction can be monitored using standard web technologies, but their focus is mainly on usability evaluation of web applications outside the lab, Carl et al. [10] record a computer user's keystrokes and eye movements which they refer to as "user activity data" in their cognitive research on natural language translators. Unlike these, our study focuses on the implications of releasing consumer-activity data back to the users. For the sake of brevity, from now on we synonymously use the term *CA data* to refer to consumer-activity data.

3. METHODOLOGY

Following the trends and background described in the previous sections, the user study presented in this paper started with the assumption that, eventually, mechanisms will be put in place by organisations to give users access to their CA data. We therefore decided to consider our own organisation, the OU, as a testbed and based on this assumption, two main questions emerged:

1. If consumer-activity data were available to users of an organisation such as the OU, what would be the benefits and use for individual users?
2. What would be the implication of deploying such a mechanism in terms of privacy risks and policies, both for the users and the organisation?

We investigated these research questions through a qualitative study which involved exposing participants to a 'prototype' tool to access and visualise their CA data collected from the OU, and then using a combination of *personal interviews*, *online questionnaires* and *focus group* to collect reactions, opinions and concerns regarding the potential uses and implications such a tool might have if properly deployed. The study itself was therefore divided into four phases - data collection, personal interviews, focus group and analysis.

3.1 The Open University

While the study could have been realised in any organisation which makes significant use of online systems, it is worth giving a more precise view of the environment in which our study was realised, its scale and relevance. The Open University is the largest university in the UK (with approx. 250,000 enrolled students per year) and is almost entirely based on open and distance learning, which means students enrol into courses upto the master's level, study at a distance and interact with the university staff (associated lecturers, course team, administrator, IT help-desk, library) mainly through online systems. Due to the size and nature of the university, the information architecture of the OU is made of a large variety of systems, most of them having a web interface and producing their own logs, which are then centralised within the IT department of the university.

3.2 Overview of the data collection process

During the data collection phase, participants were recruited through adverts on the intranets, mailing-lists and word-of-mouth. Upon their written consent, the IT team of the OU was notified with the participant's identifying details (their computer user-name), to enable them to collect (extract) the CA data from server logs of different online systems for a period of 28 days.

As a side note, it is worth mentioning that even though organisations such as the OU had been collecting user's CA data on various online systems, since there was neither any requirement nor any mandate for them to release these data to their users, significant efforts were required in making all the stakeholders (IT security department, data protection officer, ethics committee) understand the process of the user study, so that they could provide us with relevant data in a usable format.

The core of our methodology depended on the UCIAD technology platform [14] which linked and integrated heterogeneous data from several online systems within OU. Using this underlying platform, a set of GUI tools were developed on top of it in the form of a *personal analytics dashboard* where a user's activities were displayed in the form of graphical visualisations.

During the personal interview, the participants made use of personal analytics dashboard to view their past activities on various online systems of the OU. Using these visuals as a trigger, the participants were asked questions on how they felt with regards to data licensing, privacy and data protection policies which apply to such information. The personal interview was unstructured and open ended because we wanted to explore if the participant wanted to use the CA data in other creative ways not envisaged by us.

Two weeks after the interview, a short online questionnaire (with 3 questions) was sent out to all the participants of the study who had already given their personal interview, this was mainly to gauge any change in their behaviour with regards to the use of OU's online systems. The topic we were investigating had the potential to produce divergent and conflicting views, therefore we included a focus group which was designed in a way where participants could resolve their views in relation to others. To make the focus group interesting and challenging, its design was informed by the initial results we obtained from the personal interview. The participants were compensated for their time with a £30 Amazon voucher.

The qualitative data from the personal interview was analysed using Grounded Theory [11] techniques and then triangulated with the results from both the online questionnaire and focus group. The findings of the study are described and discussed later in another section.

3.3 Participants and data

The user study had 12 participants who had enrolled on a voluntary basis and were regular users of the OU's online systems. Participants broadly fell into four categories of users: post-graduate students (3), academic staff (3), academic-related staff (3) and administrative/management staff (3). The post-graduate students are PhD students who were located on the main campus of the university. The academic staff included both lecturers and researchers in different departments (i.e., Computer Science and Arts). The academic-related staff mainly consisted of technical staff who worked on IT development projects. Administrative and management staff were from different departments, including the university library and research school. This selection of participants was intended to reflect the different types of usage of the OU's online systems. However, compromises had to be made on the male to female ratio. Hence, there were 4 female participants and 8 males. The participants' age-group approximately ranged between 25 to 55 years of age.

For each participant, we collected information regarding their usage of OU websites through web server logs associated with these different systems. This required filtering these logs to keep only the data related to the participants of the study. This data collection mechanism was run over a period of four weeks (or 28 days), leading to the 12 datasets (one per participant) that included information about access and requests to OU websites. This information is encoded in a format similar to the one of Apache logs, in a text file where each line correspond to a request to a Webserver, including the following pieces of information:

> <date-time> <server> <IP of client> <user-name> <resource accessed> <response code and size> <user agent used (browser)>

This information was collected from 9 different servers, corresponding to the virtual learning environment (6 servers), the intranet, the public website and the student services website of the OU. As expected, information collected for different participants varied widely, depending on their roles. It is interesting to see for example that researchers (with no other roles) would make little use of the online systems, besides a few services provided on the intranet (e.g. expense claims, notice board, etc.) Naturally, students and associate lecturers have greater use of the virtual learning environment, while academic-related staff, especially admin staff, use a variety of online services.

3.4 Technology platform

To realise the study, a data processing and visualisation platform was required to first process and integrate the data obtained from logs into easily interpretable and exploitable CA datasets; and secondly, to create an interactive interface, i.e. the *personal analytics dashboard*, which allowed users to visualise and interact with their own CA data.

For processing and integration of log data into CA data, we reused and employed the principles and tools developed as part of the UCIAD project (see [14] for a summary) which had facilities to:

- Convert and integrate the data from their Web server log format into RDF [45], using the schema provided by the UCIAD ontologies.
- Create ontology level definitions for different types of resources and activities which are then used to process and categorise the traces of activities for the different users.
- Realise additional ad-hoc processing of the data, to improve interpretation and visualisation. For example, geographic location of the user was derived by passing the IP address to external Web services /APIs. Similarly, human readable labels for the user agents (e.g., "Chrome 8") were extracted from complex user agent strings found in the logs.
- Create a data endpoint for each of the participants based on the generated and processed RDF datasets. In this case, we used the Fuseki triple store [4], creating one separate data repository for each participant.

Once the data was processed and made available through data endpoints, the personal analytics dashboard from UCIAD provided the end-user interface to visualise and access these data. Normally, the logs from Web servers are used to produce web analytics (i.e. website usage statistics) and many users are aware of such analytical tools and interfaces associated with web analytics. We were thus motivated to design the personal analytics dashboard so that it used similar visualisation methods, but showing (instead aggregated information about visits to a website) information about the user's visits to various websites. These visualisations include information about the types of activities performed on online systems, the resources accessed, locations of the user at the time of usage, the browsers and operating systems used, the time, etc. Figure 1 shows a screenshot of the dashboard. It is important to also mention that these visualisations are interactive, as they allow the user to define filters on the activities (e.g., only show activities realised at a certain location, or at a given time) by clicking on the corresponding chart elements.

3.5 Interviews based on tool usage

The core of the user study was the personal interview conducted with each participant. The main aim of the interview was to explore the participant's reactions and views on consumer-activity data, firstly as a potential user of the system and secondly in relation to anything that they may discover which was either interesting or worrying as they used the tool. The interview started with a small introduction to the study, and initial questions about the participant's background, including their view on how they used OU's online services and their knowledge of web analytics. After this, each participant was given access to a computer running the UCIAD personal analytics dashboard, which accessed the data endpoint specifically created for them. As they used and explored the tool, they were asked to answer a number of questions related to the following topics:

- Usage of consumer-activity data ("Is there anything surprising/interesting in the data?", "Would you like to be given access to this data?")
- Data gathering issues ("Where you aware such data was being collected?", "Is there anything in what you see causing you concern?")
- Activity data policies ("Who owns the data and how should it be handled?")

The interview was conducted by two project team members: while one member asked the questions (guided/prompted by the usage and exploration of the personal analytics dashboard), the other team member took notes of the interview. Interviews generally lasted between 45 minutes and 2 hours. Each interview was audio recorded and stored in encrypted containers (to protect privacy) and later transcribed for analysis (the transcripts were anonymised before use).

3.6 Online questionnaire and focus group

An online questionnaire was sent to participants two weeks after the initial interviews in order to check whether additional thoughts and reactions would have emerged after the interviews. These questionnaires did not lead to additional insight, except to confirm that the participants had not changed either their views or their behaviour with regards to how they perceived and used OU's online systems.

The last phase of the study was the focus group. We included a focus group because all participants were from diverse backgrounds, roles and views, there was value in letting them debate their positions and viewpoints with others in a controlled setting such as a group discussion. To be effective and yet finish within the allotted time of 90 minutes, the focus group was constructed around a specific task: the group members had to collaborate with each other and contribute towards writing a business case supporting the deployment of a CA data service at the OU. This business case included two main sections: (i) benefits of CA data and (ii) obstacles to the deployment of CA data services (which indirectly pointed to privacy and security issues). During the focus group discussion, participants were asked to identify points to add to both these sections, and to react to the points made by others. Each point raised was recorded on a common document (the "draft business case"), which was being projected on a screen for all the participants to see, as a member of the UCIAD team was adding notes to it, so that it was corrected and validated collectively at the time the notes were made.

4. FINDINGS

In this section, we summarise the main results of the study, as a set of general findings, mainly regarding the benefits, privacy risks and challenges relating to potential users of CA data within an organisation such as the OU. These observations were obtained through analysing the transcripts of the recordings made during the individual interviews, together with the results of the discussions summarised from the focus group.

4.1 Potential benefits in releasing CA data

As previously described, CA data is a by-product of the user's interaction with the organisation's online systems.

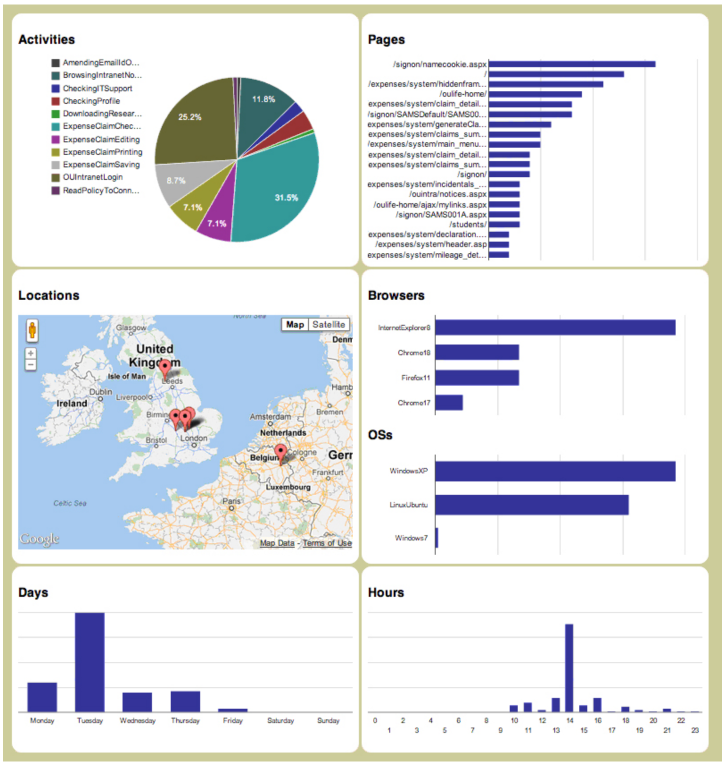

Figure 1: Screenshot of the UCIAD personal analytics dashboard for a user.

Therefore, one of the fundamental question the user study aimed to answer is whether users will be sufficiently interested to request a copy of the data and put it to any use. It was interesting to note that the participants came up with several innovative suggestions as to how they may be able to use their activity data to better themselves:

Self reflection: Some participants of the study saw value in simply being able to reflect on what they do daily, it was more of a self assurance tool to indicate if they were on track with their actions or not: *"yeah...it is for self reflection, it can show you if you are on track...that you are not deluding yourself as to what you are actually working on" [P3]*. This notion of self-reflection on one's own activities is generally related to the one of lifelogging [35] (see in particular [13] that discusses this idea in relation to web interactions).

Self improvement: Another interesting benefit for CA data the participants suggested was to improve their use of online resources. This is very closely related to self-reflection and it is an outcome of what they see as "inefficiencies" in their use of the resources. For example, users find it frustrating not to have bookmarked certain pages they use quite frequently and to have to search for them: *"I kind of realised this is not very effective and every time I need that information about the poster...[I want to] bookmark it, you know...to myself and then I never do...but when you look at something like this...you go like...put some effort and bookmark it" [P5]*.

Trace anomalies: Very similar to how application logs have long served the purpose for analyses and tracing error conditions in the software, users find the CA data useful to trace anomalies in their work pattern. For example, they can trace back their online activities to specific websites whose security had been compromised: *"I can think of an impact around sort of data security and personal security and it being able to flag up... doing something on a browser that might not be so secure as you might want or looking at websites that are actually security risks...to actually be able to go back and track back to say actually did I use that website at that particular time...thats the type of thing that will be useful" [P7]*.

Promote transparency: In the context of the organisation (the Open University), participants felt that the very act of releasing the CA data to the users would significantly improve the organisation's reputation: *"if you have a view of the data who is holding, I think it makes...it reassures the users I suppose, this is what the OU [organisation] knows, this is what you know...that in a way quite equal" [P10]*. Participants also feel that CA data can provide transparency on a personal level, for example, the users are able to justify or deny their actions using the CA data as evidence: *"as long as I am doing my job I don't think its a problem...but I suppose...I am trying to think...my time in Berlin is accounted for so I am quite happy that people know that I was in Berlin" [P4]* and *"I am not...you know...I am not doing anything that I wouldn't be able to stand up and defend in court if I had to" [P9]*.

Although these benetifts may not be extraordinary in themselves, it shows that users when given an opportunity to access their CA data, they are likely to find innovative and creative means to benefit themselves. Given that users are likely to exploit their CA data, we focus on the potential privacy and security risks, both the users and the organisation, are likely to face especially with regards to key questions such as: (1) In what ways will the user's privacy be harmed by this data? (e.g. in case of data leakage); (2) Can the organisation suffer any privacy harm? (e.g. if the user voluntarily shares the data on social media) and; (3) What must be done to avoid harms from (1) and (2)?

4.2 Privacy risks: users and organisation

First, we describe the privacy risks. While these privacy risks may not be completely new, it does underline and confirm that even in a trusted producer-consumer relationship (as in this case), such privacy threats exist, and it must be addressed.

Image distortion: Distortion seems to be a major concern. Distortion refers to disseminating false or misleading information about an individual [40]. The user study showed three contributors to distortion: incompleteness, inaccuracy and incorrectness of data. In the interviews, one of the most commonly expressed concern related to the *incompleteness* of CA data. Some participants felt the indicated usage were skewed and incorrect because it excluded data from several other online sources, and the resulting incomplete picture presented them in negative light. For example, one user had indicated the missing CA data could be misunderstood as them "not working", when in fact their activities had been captured elsewhere which was not included in the study: *"I would like to be able to put in my wider browsing history and quite like as well to pull in my calendar schedule of meetings stuff like that and just see how much time I actually spent because like I said there times when I in a meeting or out somewhere else...so my use of particular systems or web in general may not appear so intense as perhaps it is... so I quite like to see that" [P9].* The users were worried such incorrect inferences made from incomplete data could hurt their reputation. The second contributor to the distortion threat was the *inaccuracy of data*, a common example related to the location of the user displayed on the map for each activity. The location of the user was derived from the IP address of the computer used in the interaction, but the location derived through this method was often misleading because the service which interpreted the IP address was grossly inaccurate in specific cases. In one particular instance, the location pointed to another city the user hadn't visited before, only adding to the fear that, without the user's input, the inferences could be wrong and harmful. The third contributing factor to the distortion threat is *incorrect* metadata. More precisely it refers to the incorrect classification and categorisation used in producing the analytics. When there is a mismatch between how the user and the system have classified an activity, it can have negative implications for the user: *"it [category/classification of activity] would have to be interpreted very carefully. So obviously it might be somebody, some social scientist who might be researching pornographic sites and its impact on society" [P5]*; in this case the social scientist could be mistaken for someone watching and accessing adult material during office hours.

Unwanted disclosure: Disclosure involves the revelation of an individual's true information to others impacting her reputation [40]. Even if the data was distorted, its impact is limited when the data is not shared with others: *"it's fine because I am looking at it [data] and I know how to interpret it, if somebody else looked at this...like you didn't know why I spent so much time on the eBook thing...yeah, I don't like the idea that people can simply look at it and interpret it in the way it wasn't correct" [P3].* It's understandable that users are worried that the incorrect inferences could be viewed by individuals within the organisation or elsewhere who can make decisions that are detrimental to the user's interest *"it can be quite misleading judging from this...this isn't my work, this isn't in my working day and if people are going to be basically making decisions based on this...yeah... it makes me nervous" [P4].* The point here is that the users understand the context of their work and their activities and expect to be able to interpret them. The underlying concern however is that if this data is accessible to others (within the organisation or outside), they might misjudge the user's reputation based on the partial data.

Re-identification: Identification refers to linking information to particular individuals [40]. Whether the inferences are accurate or not, generally users don't seem to be overtly concerned if the CA data is anonymised, but it does cause concern if they can be identified or the inferences can be associated with them on an individual basis: *"I don't mind if it is aggregating vast data...its perfectly sensible in my book but its when they are drilling down on into individual usage or even individual departments usage then I'll get a little bit more nervous I think." [P4].* Although the data is anonymised, there is always a possibility that the users might be identified through triangulating with other datasets.

Opaqueness of passive data collection: Exclusion is failure to provide individuals with notice and input about their records [40]. Currently, organisations provide privacy notices when personal data is being collected but it is usually in general terms and, as such, they are not obliged to specifically mention the nature of CA data their online systems generate. Although organisations may be highly trusted and may even be seen in favourable light, there is an implicit expectation that the organisation would be as open and transparent as possible: *"in principle I don't have a problem but I guess just thinking about it...yeah...who is collecting the information for what purpose...I think I'd want to be made aware of that...before the information is collected not afterwards...just nothing...apart from being kept in the picture" [P4].* At this stage it is not very clear if the users will change their behaviour after being made aware of data collected by their online interactions, but nevertheless they would like to be notified.

Insecurity from compromised infrastructure: Insecurity is failure to protect personal data [40]. Assuming organisations decide to make the CA data available for the users to download, this would require special secure channels to be opened. However, it still entails a risk where malicious attackers may obtain a copy of the data: *"yes, it would nice to have access to it, I'll be interested but if you set-up a system like that you open a gate, no matter how secure it is...there's a risk people who should not see...I mean other people will be trying, someone might break-in or whatever" [P2].* Organisations spend a lot of their resources in securing their infrastructure, the risk may be lower when compared to how users will be able to protect their copy of CA data once it has be retrieved from the organisation's system. In this respect, the user is far more vulnerable for attacks and data leakages through loss of storage devices or laptops.

Breach of personal and organisation's confidentiality: Breach of confidentiality is unauthorised revelation of confidential information resulting in loss of trust [40]. The CA data not only contains references to the user but also holds information about the organisation, which can be exploited by the user (e.g. disgruntled employee) to damage the organisation's reputation: *"what you wouldn't want to happen is say if you were the employer that we make this available to people and suddenly they find a way to use to the detriment of the organisation." [P8]*, especially in cases where the users decide to upload the content on to social networking services. This would breach the trust the organisation has on its employees.

3rd party exploitation of data: Secondary use is the use of data for purposes unrelated to the purposes for which it is collected [40]. During the study, many of the users were not aware of their data being collected by external third party services such as Google analytics [17]. While Google Analytics aggregates data from several users of particular online systems to provide usage statistics, they also exploit the data they gather for their own purposes, to create user profiles which can help them in targeting advertisements. In the study, not many of the users were aware and wanted Google to collect their data, although some of them, without being fully aware of the consequences, had added Google Analytics to their project web page/websites. Third party use of CA data can only be external, since users cannot be expected to be able to develop tools to process the raw CA data and will be dependent on third party services to analyse these data. In such a scenario, the CA data which contains information provided by the organisation about the user's activities might be used for purposes not intended by the user or the organisation. This therefore constitutes a potential privacy risk for both of them.

4.3 Privacy policy models: ownership and licencing

As shown in Table 1, the seven potential privacy risks that apply for both the users and the organisation depending on where the CA data is located. By default, the CA data is with the organisation: this is the state when the data has been created but not downloaded by the user. When the user downloads the data, it is considered to be in the user's domain.

At a higher level, organisational privacy policies govern how the data is collected, processed and to what use it will be put to. It relates to ownership and licensing of data, here *ownership* is concerned with controlling who can change the access permissions and usage rights on the data. The owner who holds the ownership rights may allow other entities to access and use the data by issuing a *license* (i.e. terms and conditions for use). As shown in Figure 4.3, the CA data may reside in the organisation or the user's domain (or both) and privacy risks apply differently depending on who has control over the data. For an effective access control policy, the access rights and ownership rights have to be clearly defined. However, in the study we found that users have no clear consensus on ownership rights that should be applied to CA data. Here we briefly describe the three potential models participants relate to:

(1) The organisation owns the data: Currently, this is seen as the default model. The organisation create the data, therefore they retain the rights and exploit the data as their own asset, and some user's acknowledge and support this position: *"I think the OU owns it...its the University's data, its information about me but University compiled it and they must own it...as an individual I don't own the data, I am in that camp...some people may say they generated the data, no [they] didn't, the University happened to collect the data of my usage, I didn't create the data" [P8].* Since the organisations are constrained by the Data Protection Act which specifies under what conditions the data can be used without having to obtain consent from the user, delivering CA data would effectively mean issuing a license to the user to not only access their own data but also use it in ways they deem fit. The CA data not only contains details about the user but also of the organisation. Assuming organisations redact the CA data before releasing it, there is always the risk that some confidential information can be triangulated through other means. It is therefore in the interest of the organisation to impose restricted licensing conditions that make it difficult for the CA data to be aggregated and used in ways that are determental to its reputation. This of course raises an important question - *what type of restrictions should the organisation be allowed to impose on the released CA data, given that users will want to exploit its benefits?* The majority of study participants indicated that they would expect the organisation to impose strong limitations on the use of the data (e.g., it could not be shared or it could only be accessed through the tools provided by the organisation).

(2) The user owns the data: The second model represents a radical shift from the default model described above. In this model, the participant's suggestion is for the users to have full rights over the CA data once it is transferred to the individual: *"if the data is given back to me then I'd expect I am the owner of it...that the people who had originally collected that data have passed the ownership to me to do what I like" [P7].* The participants reason that, since the CA data represents personal information, it is only natural for the individual to be in control of its use. One major implication of this approach is that the organisations by default will not have the right to either access or exploit the CA data in any form, unless it has been explicitly licensed by the user. Therefore, the important question here is - *what type of restrictions should the user be allowed to impose on the organisation's use of CA data?* Allowing users to impose conditions could make it more complex for organisations to genuinely use the CA data to improve their online services and, as a consequence, make the collection of CA data irrelevant and counter-productive for organisations.

Another reason to transfer the ownership to the user is that the Data Protection Act indicates that personal data might not be kept by an organisation for longer than necessary for the intended purpose. Since organisations keep CA data for a short period (at the OU, the data is kept for 7 consecutive days) and then deleted, one possible approach would therefore be to transfer the data together with its associated rights to the user after this retention period. From a practical perspective, this would also represent a convenient arrangement, as users could simply 'opt-in' for their CA data to be sent to them at the time it is deleted from the organisation's systems. Of course, this would still imply that the organisation would lose control once the data has been transferred to the user, and "non detrimental use/confidentiality" clauses might be attached to the delivery of CA data.

(3) Shared data ownership: The third suggestion is for a hybrid approach where both the organisation and the user are considered as co-owners of CA data. This means that both will have full access to the data, but would be

Privacy Risk Matrix	Breach of Conf.	Exclusion	Distortion	Disclosure	Identification	Insecurity	Secondary use
CA data → Organisation domain	x	x	x	x	x	x	x
CA data → User domain	⊗			x		⊗	⊗

→ **Location of data** x **- Privacy Risks to User** ⊗ **- Privacy Risks to Org. + User**

Table 1: Matrix of privacy risks applicable in user and organisation domains.

prevented from using them in ways that could be detrimental to each other. Study participants proposed this to be an appropriate model: *"it is shared data almost isn't it, like pay scales and salaries and things...that is OU data as well...as to what they pay their staff, so its not just mine to do whatever I want...its like a shared ownership" [P3].* The legal implementation of such an arrangement could be rather complex, or on the contrary it might be simpler than anticipated. For example, the organisation can comply with the Data Protection Act just as they do when the CA data in their domain, while at the same time, the users can be bound to confidentiality agreements with the organisation, thus preventing them from widely distributing these data. In any case, new terms and conditions will have to be drawn, so the important questions are: *what are these joint terms and conditions? Will there be any flexibility in negotiating them? How will conflicts be resolved?*

We found that there was no clear consensus among the participants on any of these ownership and licensing models described above. We also found that participants were not in position to articulate and elaborate on the terms and conditions that might entail in each of the three models so that both, users and organisations, can mutually benefit from the use of CA data.

5. DISCUSSION

Although the study had a sample size of 12 participants, we do not claim this to be representative of all CA data users. The aim of our study was to explore and highlight the potential impact releasing of CA data will have on both the users and the organisations. The study, showed that if given an opportunity, users can come up with ingenious ideas to exploit and benefit from CA data. We also identified some of the potential privacy risks which are applicable to both the users and the organisation. Although the discovered privacy risks are not new, the study showed these risks are dormant even in datasets generally not visible to the users. More importantly the privacy risks relates to both the user and the organisation and here we briefly discuss the impact of these findings with respect to two topics:

Legistlation and principles: The Data Protection Act which controls how organisations, businesses and the government process data, applies primarily to personal data. This then throws open an important question: *should CA data be considered as personal data?* In our study, CA data contained traces of the user's identity, for example, the online system's user-name but this in itself was insufficient to link to any individual within the organisation (without having access to other directory systems within the OU). Everything in the Data Protection Act pertaining to data access rights of both the users and the organisations then hinges over this classification. Considering CA data as personal data can have far reaching implications under the Data Protection Act: (a) users will be allowed to demand a copy of their data unlike the current situation where organisations are encouraged to release CA data voluntarily, (b) organisations will be forced to make explicit how they process the data and for what purposes it is being used, and (c) organisations will be unable to allow 3rd party use without explicit consent from the user. Current ambiguity attached to the nature of CA data means organisations are not obliged to provide notice and obtain the user's consent regarding (b) and (c). For other implications, the Data Protection Act in its current form appears to be inadequate. For example, the CA data, in addition to containing traces of user's identity, also contains confidential information about the organisation's internal structure, therefore the question arises: *what data access rights do organisations have after the data has been transfered to the user's domain?*, which of course requires further debate and legislation.

Privacy enhancing techonologies: Following the legislative implications, here we assess the impact on privacy enhancing technologies which might be required to support the release and use of CA data, for example, the questions to consider are: *How to provide notice and consent for the CA data being collected? To manage access rights, how can both the user and the organisation jointly author and enforce privacy policies? When storing CA data, how to secure it both on the organisation's server and on the user's personal computer?*, and *To what extent can the CA data be anonymised so that it is still useful?* We breifly look at these implications:

(a) User awareness and consent: It would be a regressive step to simply provide notice for CA data collection in the form of a lengthy legal document when previous research has shown that such policies are difficult for users to read [21] and that this format may not be effective [28]. Specifying compact machine-readable privacy policies in W3C's P3P [44] and APPEL [43] is not an option either, given the poor uptake it has had over the years. Moreover, contrary to assumptions made by P3P implementations, the study participants have indicated that they do not have an *"adversarial relationship"* with the organisation and the considered online systems are critical to accomplishing their daily tasks. One option could be to explore the use of privacy icons [24] and "nutrition labels" [23] to indicate the type of data that is being collected by the online system and also provide options for the user to opt-out if required.

(b) Privacy policy languages and tools: Organisations have security and privacy experts to define privacy policies in as many complex languages (e.g., XACL [18], XACML [32], SAML [33] etc.), but users on the other hand may be novices and non-experts who may not be able to easily express their

privacy policies in an unambiguous way. Thus, there is an implicit requirement for innovative user interfaces to front privacy policy editors to improve usability. Since both the user and the organisation would have to exchange privacy policies (depending on the ownership model), they will need to agree upon a common (high-level) privacy language. For example, it could be based on EPAL [5], DPAL [9] or perphaps invent a new language tailored just for this purpose. Another concern relates to policy enforcement in heterogeneous domains, first in the organisation where the CA data is created and initially used and then in the user domain when the CA data has been transferred. Current approaches such as E-P3P [6] and 'super sticky' policies [8] which propogate policies in multiple domains are aimed at enterprises and organisations are not particularly suited in a "non-enterprise" setting (when the user downloads the CA data at home or on their personal desktop/laptop).
(c) Data storage and transmission: Currently, novice users have little or no knowledge of using encryption for storing data and therefore cannot be expected to use such techniques to keep the CA data safe from accidental loss. Even after several years, the challenges identified in the initial report [46] on making encryption more usable remain open [39]. New architectural approaches to securing data in "personal containers" [30, 36] and "personal data vaults" (PDV) [31], look more promising. Such approaches not only support data streamed from heterogeneous sources but also secure communication channels and provide user defined access control. However, in their current form these architectures may not support the type of ownership and licensing models described earlier, mainly because such models demand privacy policies be expressed as "joint" conditions between the user and the organisation.
(d) Anonymisation tools: The CA data contains identifiable details of both the user (e.g. the user name) and the organisation (e.g. IP address) which can lead to the privacy risks of identification. While anonymisation techniques such as k-anonymity [41] claim to make it difficult for attackers to identify a user through linking of multiple datasets, in practice, implementations of such techniques have their own complexities [29]. Achieving an "optimum" level of anonymisation will be difficult as anonymisation conflicts with the data aggregation and mining techniques which the user and the organisation may want to employ in order to exploit the benefits of CA data.

6. CONCLUSION

With the growing trend of consumer-activity data being given back to users, it is important to understand and anticipate both, the benefits these users might gain from it, as well as, crucially, the privacy implications it might entail. We investigated this in the context of an large educational institution by conducting an explorative qualitative study. For the study, consumer-activity data was gathered from several online systems which the users interacted with to accomplish their task. Using an analytical dashboard as trigger, the study explored the participants' reactions to how they perceived their own data and the potential benefits for them if they were given access to such data. In the process, the study uncovered 7 privacy risks for the users, out of which 3 threats were also applicable to the organisation. This paper also highlights the data ownership and licensing models relevant to both the users and organisations, and finally, discusses and provides pointers to challenges that must be addressed before consumer-activity data gets released to users.

7. ACKNOWLEDGMENTS

The study presented in this paper was partly achieved thanks to funding from JISC for the UCIAD (User Centric Integration of Activity Data) project – see *http://uciad.info* for details.

8. REFERENCES

[1] Explore Your Activity Log (Facebook). www.facebook.com/help/activitylog. [Accessed: 5th March 2013].

[2] The Data Liberation Front. www.dataliberation.org. [Accessed: 5th March 2013].

[3] UK Government, Midata: access and control your personal data. www.bis.gov.uk/policies/consumer-issues/personal-data. [Accessed: 5th March 2013].

[4] Apache Jena. Fuseki: serving RDF data over HTTP. http://jena.apache.org/documentation/serving_data/index.html. [Accessed: 6th March 2013].

[5] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter. Enterprise Privacy Authorization Language (EPAL). www.w3.org/Submission/2003/SUBM-EPAL-20031110/. [Accessed: 6th March 2013].

[6] P. Ashley, S. Hada, G. Karjoth, and M. Schunter. E-P3P privacy policies and privacy authorization. In *Proc. of 2002 ACM workshop on Privacy in the Electronic Society*, pages 103–109, Washington DC, 2002. ACM.

[7] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction.

[8] S. Bandhakavi, C. C. Zhang, and M. Winslett. Super-sticky and declassifiable release policies for flexible information dissemination control. 51.

[9] A. Barth, J. C. Mitchell, and J. Rosenstein. Conflict and combination in privacy policy languages. In *Proc. of 2004 ACM workshop on Privacy in the electronic society*, pages 45–46, Washington DC, 2004. ACM.

[10] A. L. J. Carl, Michael and K. T. Jensen. Studying human translation behavior with user-activity data.

[11] J. Corbin and A. Strauss. *Basics of Qualitative Research, Techniques and Procedures for Developing Grounded Theory.* Sage Publications, 2008.

[12] M. d'Aquin, S. Elahi, and E. Motta. Semantic monitoring of personal Web activity to support the management of trust and privacy. In *SPOT 2010: 2nd Workshop on Trust and Privacy on the Social and Semantic Web*, Heraklion, Greece, 2010.

[13] M. d'Aquin, S. Elahi, and E. Motta. Semantic monitoring of personal Web activity to support the management of trust and privacy. In *Proc. of WebSci10: Extending the Frontiers of Society On-Line*, Raleigh NC, USA, 2010.

[14] M. d'Aquin, S. Elahi, and E. Motta. Semantic Technologies to Support the User-Centric Analysis of

Activity Data. In *Social Data on the Web Workshop, SDoW 2011 at ISWC 2011*, 2011.

[15] S. Elahi, M. d'Aquin, and E. Motta. Who Wants a Piece of Me? Reconstructing a User Profile from Personal Web Activity Logs. In *Intl. ESWC Workshop on Linking of User Profiles and Applications in the Social Semantic Web*, 2010.

[16] Federal Trade Commission, USA. Fair Information Practice Principles. www.ftc.gov/reports/privacy3/fairinfo.shtm. [Accessed: 5th March 2013].

[17] Google-Inc. Google Analytics. www.google.com/analytics. [Accessed: 6th March 2013].

[18] S. Hada and M. Kudo. XML Access Control Language: Provisional Authorization for XML Documents. www.research.ibm.com/trl/projects/xml/xacl/xacl-spec.html. [Accessed: 6th March 2013].

[19] L. Hasan, A. Morris, and S. Probets. Using Google Analytics to Evaluate the Usability of E-Commerce Sites. In M. Kurosu, editor, *Human Centered Design*, volume 5619 of *Lecture Notes in Computer Science*, pages 697–706. Springer Berlin Heidelberg, 2009.

[20] Information Commissioner's Office(ICO), UK. Key definitions of the Data Protection Act. http://tinyurl.com/8f5wlbp. [Accessed: 5th March 2013].

[21] C. Jensen and C. Potts. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, pages 471–478, Vienna, Austria, 2004. ACM.

[22] JISC, UK. Activity Data. http://tinyurl.com/6lcb5xw. [Accessed: 5th March 2013].

[23] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder. A "nutrition label" for privacy. In *Proc. of 5th Symposium on Usable Privacy and Security*, pages 1–12, Mountain View, California, 2009. ACM.

[24] KnowPrivacy. Policy Coding Methodology. http://knowprivacy.org/policies_methodology.html. [Accessed: 6th March 2013].

[25] B. Krishnamurthy and C. E. Wills. Generating a privacy footprint on the internet. In *Proc. of 6th ACM SIGCOMM conf. on Internet measurement*, pages 65–70, Rio de Janeriro, Brazil, 2006. ACM.

[26] C. Mancini, Y. Rogers, K. Thomas, A. N. Joinson, B. A. Price, A. K. Bandara, L. Jedrzejczyk, and B. Nuseibeh. In the Best Families: Tracking and Relationships. In *Proc. of 29th Intl. Conf. on Human Factors in Computing Systems, ACM CHI 2011.* ACM Press, 2011.

[27] C. Mancini, K. Thomas, Y. Rogers, B. A. Price, L. Jedrzejczyk, A. K. Bandara, A. N. Joinson, and B. Nuseibeh. From spaces to places: emerging contexts in mobile privacy. In *Proc. of 11th Intl. conf. on Ubiquitous computing*, pages 1–10, Orlando,USA, 2009. ACM.

[28] A. M. Mcdonald, R. W. Reeder, P. G. Kelley, and L. F. Cranor. A Comparative Study of Online Privacy Policies and Formats. In *Proc. of the 9th Intl. Symposium on Privacy Enhancing Technologies*, pages 37–55, Seattle, WA, 2009. Springer-Verlag.

[29] A. Meyerson and R. Williams. On the complexity of optimal K-anonymity. In *Proc. of 23rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, Paris, France, 2004. ACM.

[30] R. Mortier, C. Greenhalgh, D. McAuley, A. Spence, A. Madhavapeddy, J. Crowcroft, and S. Hand. The Personal Container, or Your Life in Bits. *Digital Futures' 10*, pages 11–12, 2010.

[31] M. Mun, S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen, and R. Govindan. Personal data vaults: a locus of control for personal data streams. In *Proc. of the 6th Intl. Conf.*, page 17. ACM, 2010.

[32] OASIS. eXtensible Access Control Markup Language (XACML). http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.pdf. [Accessed: 6th March 2013].

[33] OASIS. Security Assertion Markup Language(SAML). www.oasis-open.org/standards#samlv2.0. [Accessed: 6th March 2013].

[34] OECD. OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. http://tinyurl.com/bgojzhu. [Accessed: 5th March 2013].

[35] K. O'Hara, M. Tuffield, and N. Shadbolt. Lifelogging: Privacy and empowerment with memories for life. *Identity in the Information Society*, 1(1):155–172, 2008.

[36] PersonalContainers. Personal containers. http://perscon.net/. [Accessed: 6th March 2013].

[37] B. Plaza. Monitoring web traffic source effectiveness with Google Analytics: An experiment with time series. *Aslib Proc.*, 61(5):474–482, 2009.

[38] B. Plaza. Google Analytics for measuring website performance. *Tourism Management*, 32(3):477–481, 2011.

[39] S. Sheng, L. Broderick, C. Koranda, and J. Hyland. Why Johnny still can't encrypt: evaluating the usability of email encryption software. In *Symposium On Usable Privacy and Security*, 2006.

[40] D. J. Solove. *Understanding Privacy.* Harvard University Press, London, 2008.

[41] L. Sweeney. k-anonymity: A model for protecting privacy. *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[42] S. J. Turner. Website Statistics 2.0: Using Google Analytics to Measure Library Website Effectiveness. *Technical Services Quarterly*, 27(3):261–278, 2010.

[43] W3C. A P3P Preference Exchange Language 1.0. www.w3.org/TR/P3P-preferences. [Accessed: 6th March 2013].

[44] W3C. Platform for Privacy Preferences (P3P). www.w3.org/P3P. [Accessed: 6th March 2013].

[45] W3C. Resource Description Framework (RDF). www.w3.org/RDF. [Accessed: 6th March 2013].

[46] A. Whitten and J. D. Tygar. Why Johnny can't encrypt: a usability evaluation of PGP 5.0. In *Proc. of 8th Conf. on USENIX Security Symposium*, pages 14–142, Washington DC, 1999. USENIX Association.