Consumer Activity Data: Usages and Challenges

A study from the *UCIAD* project

Mathieu d'Aquin and Keerthi Thomas
Knowledge Media Institute
The Open University

Motivation and Background

Interacting online with various organisations (eCommerce, employer, etc.) is nowadays unavoidable. As in the offline world, such online interactions require the exchange of information between the different parties involved. When looking at personal data however, it is natural for such an exchange to be unbalanced: the individual interacting with an organisation will in most cases be the producer of information, and the primary topic of the data exchanged, but these data will mostly be collected and consumed by the organisation, outside the individual's control. The rational here is of course that such data is of economic value to the organisation, which can aggregate it and make use of it to provide better services to the user. There is however a current trend both in academia and the industry taking as a starting point that such personal data is also of value to the user, and that putting them out of his/her control might have problematic implications. A proposed solution here is to somehow invert the balance towards the users (or consumers, in reference mostly to eCommerce organisations, leading to the term "consumer data"), by providing them with greater access to and control over personal data collected out of their interactions with an organisation. It is worth noticing here that this idea is actually at the basis of the *Data Protection Act*, which is providing a legal framework to enforce transparency and user control over the collection of personal data. The current focus on consumer data is however both more concerned with the technological aspects of ensuring transparency, and proposing to go a step further by enabling transparency "by default" rather than on request.

Personal data in such discussions can take many different forms. The most common and trivial one being information entered by the users that is strongly associated with their identity: name, email address, telephone number, etc. A second level of personal data concerns the social connections shared between individuals, as well as with organisations (friends, married, customer, student, etc.) What we are concerned with here is a further level of personal information (in terms of how indirectly it relates to the user's identity) that is currently being broadly under-investigated from the point of view of transparency: activity data. In short (more details are given below), we consider activity data as any type of information that is being generated as a side effect of an individual interacting with an organisation, system or set of resources. The basic hypothesis underlying the study described in this report is that, while rarely included in the discussions around consumer data, as for any kind of personal information, activity data are valuable to users and it is therefore worth investigating the mechanisms, issues and challenges by which they could be made more transparent and more controllable.

What is Activity Data

As briefly mentioned above, we consider activity data as being all of the information that is generated as a side effect of an individual interacting with an organisation, a system or a set of resources, and that concerns the traces of these interactions. We can see mainly two levels of activity data: 1- related to high level, application specific transactions with the organisation, system or resources, or 2- related to the low level traces of interactions that are generated out of using and accessing the corresponding websites.

The first category concerns transactions such as "buying a product" or "commenting on an article" which are very high level and specific to the particular environment in which they are realised. While many online systems now provide to their users views on such transactions (and they tend to be in the scope of consumer data), there are rarely exploited beyond the simple purpose of keeping records.

What we are more interested in here is the second category of lower level (and by extension richer and more granular) traces of interactions. These concern the records, or logs, that are collected out of accessing individual resources, webpages or features of online systems. The reasons why these are especially interesting is that they reflect different perspectives of the user's behaviour in interacting with online systems at a very granular level, with the associated data being generally very dense. Such data are mostly commonly used by organisations as a way of monitoring the functioning of their systems and websites, as well as to produce analytics of the aggregated behaviour of users on these systems and websites. Activity data in this sense is however rarely considered as part of consumer data, for a number of reasons we will detail later in this report. It is actually unclear whether such level of information is even considered a type of personal data, even if it is clearly associated with identifiable individuals.

What is Consumer Data

In general terms, Consumer Data refers to information that relates to particular users of an organisation, system or set of resources. More specifically in the context of this report, we use consumer data as the set of approaches and mechanisms used to enable access and control from individuals to the personal data generated as a result of their use of particular websites and online systems. The proper handling of consumer data is envisioned as being of critical importance both for the economy, and for the benefits of individuals, as illustrated by the *UK midata project*, which goal is to "empower individuals with their personal data" (see *midata empowerment*). The basic principles and requirements for such realisation of consumer data management is summarised by the midata project in "*Midata Consumer Data principles*" referring in particular to notions of data reusability and data standards. It is expected for such a approach of consumer data to actively participate in areas related to "Vendor Relationship Management", as it applies similar ideas of reversing the balance towards consumers in their interactions with organisations (especially, commercial ones). Following similar trends, many projects (especially in the technology sectors) have emerged with the aim to develop the technological support for consumer data management, including the Google "*Data Liberation Front*", the *Danube project* or, of course, UCIAD (*http://uciad.info*).

About the UCIAD project

The *UCIAD* project (User Centric Integration of Activity Data) is a JISC-funded project investigating the issues related to providing back to users access and control over their activity data, as consumer data. The first phase of the project was dedicated to tackling the technological issues related to collecting, managing and distributing activity data centred on a particular user. One of the main issues was the need to integrate large amounts of data from different systems in a structured way, which can be easily interpreted and manipulated by user-facing tools. The approach taken was based on the use of semantic technologies (*RDF*, *SPARQL* and *OWL* Ontologies) to provide a common model into which logs from various systems could be imported,

and reasoned upon. The results of this phase are summarised in the article "*Semantic Technologies to Support the User-Centric Analysis of Activity Data*" [1].

In the second phase of the project, the idea was to use the base technological infrastructure developed in the initial part of the project to investigate the use cases and organisational issues implied by consumer activity data, in the environment of the Open University. In other terms, based on a user study of the application of UCIAD technologies, the goal was to investigate the two following questions:

1. If consumer activity data were available to users of an organisation such as The Open University, what would be the scenarios making such a mechanism useful?
2. What would be the implication of deploying such a mechanism in terms of organisational policies?

The goal of this report is to provide an overview of the approach taken to investigate these questions, and of the results obtained.

Methodology

The intent here is to investigate the two questions above, based on a concrete user study. The main difficulty however is that the two questions are formulated in such a way that they rely on an assumption which is not valid nowadays: that activity data would be provided back to users as consumer data. To tackle this issue, at the core of our methodology is a prototype set of tools for the user centric integration of activity data, where the data is taken from various online systems at the Open University. In other terms, the core of our methodology can be summarized in the following steps (which will be detailed further in the next sections):

1. To employ the technology developed in UCIAD to develop a *personal analytics dashboard*, showing users information about their activities on Open University online systems
2. To collect data from volunteer participants to feed the technological platform developed within UCIAD
3. To conduct individual interviews based on the exposure of users to the UCIAD personal analytics dashboard
4. To complement these interviews with an online questionnaire as well as a group discussion, where views and ideas on consumer activity data could be exchanged

The general idea of this study was to collect thoughts and reactions from participants (as prospective users of a consumer activity data service), in order to identify prominent use cases and potential challenges in this area. The results are summarized later in this report. In the following section, we give more details of the specific settings of the study at different stages.

Participants and Data

The study was realized with 12 participants enrolled on a voluntary basis amongst the users of the Open University's online systems. Participants fall into different categories of users (students and/or staff, including associate lecturers and/or academic related staff and/or researchers). This was intended to reflect the different types of usage of Open University's online systems.

For each participants, we collected information regarding their usage of Open University website 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 ran over a period of four weeks (covering more or less the month of April 2012), leading the 12 datasets (one per participant) that included information about access and requests to Open University websites. Such 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> <username> <resource accessed> <response code / 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 Open University.

As expected, information collected for different participants vary 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.

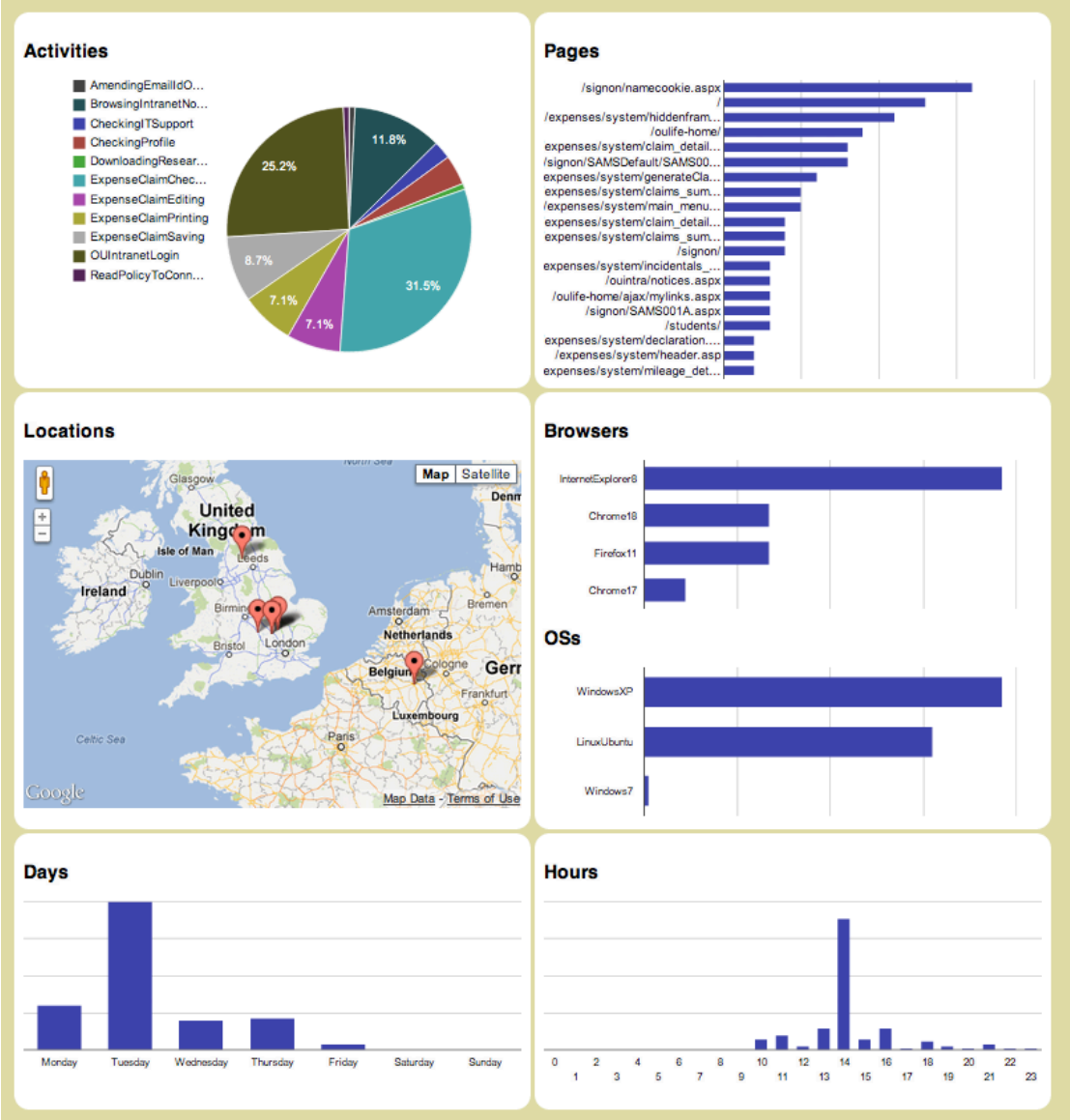
Technological Platform

The goal of the UCIAD technological platform is twofold: 1- to process and integrate the data obtained from logs into easily interpretable and exploitable, user-centric datasets of activity data; and 2- to create an interactive interface, the UCIAD personal analytics dashboard, to allow users to interact with their own activity data.

Regarding the processing and integration of log data into activity data, we reused and employed the principles and tools developed as part of the first phase of the UCIAD project (see the paper [1] for a summary). This involved in particular building tools to:

1. Convert and integrate the data from their log format into RDF, following the schema provided by the *UCIAD ontologies*. The *UCIAD Parderer* tool was used for this.
2. Create ontology level definitions of types of resources and activities to enrich the initial data, to then apply ontology reasoning as a way to classify traces of activities according to these ontological definitions (following the principles described in [1]).
3. Realise additional *ad-hoc* processing of the data, to improve interpretability. This included in particular deriving the location of the user from their IP (using a dedicated online API), deriving human readable labels for the user agent (e.g., “Chrome 8”) from the complex user agent strings included in the logs, deriving general date/time information from the timestamp included in the logs (e.g. day of the week), etc.
4. Create a data endpoint for each of the participants based on the generated, processed RDF datasets. In this case, we used the *Fuseki* triple store, creating one separate data repository for each participant.

Once the data processed and made available through data endpoints, a crucial aspect of the UCIAD technological platform was the end-user interface to visualize and access this data. The data collected would most commonly be used for the purpose of web analytics and many users are familiar with the idea and interfaces associated to web analytics. We therefore designed the interface to the data as a personal, user-centric analytics dashboard which, rather than showing aggregated information about visits to a website, displays to a user information about their visits to various websites (including classifications of activities, resources accessed, location, browsers and systems, time, etc. – see screenshot below).



Screenshot of the UCIAD personal analytics dashboard for a particular user.

On a side note, it is worth mentioning that this interface is an interesting system by itself, from a technological point of view. Indeed, it is developed using Javascript and accessing the relevant data endpoints directly from the client's browser, through the *SPARQL 1.1 query language and protocol*. It shows charts and visualization created dynamically using the Google *Chart* and *Maps* APIs and is highly interactive, as any

value/visualization element displayed can be clicked to produce filters on the data (e.g., restricting the display to activities realized “while at home on a Sunday” for example). The Javascript source code of the UCIAD personal analytics dashboard is also *available online*, under an open license.

Interviews Based on Tool Usage

The core of the user study was a series of interviews realised with each individual participant. The base idea for these interviews was to collect direct reactions and opinions about consumer activity data from potential users, as they discovered and tried the tools. Therefore after a small introduction to the study, and initial questions about their background, their use of Open University online services and their knowledge of the area of web analytics, each participant was given access to a computer running the UCIAD personal analytics dashboard on their corresponding data endpoint. 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 that could cause concern?)
- Activity data policies (In what form do you think this data should be available? What should be the arrangement in terms of data ownership?)

The answers to these questions were recorded in the form of a conversation between the participant and the members of the team conducting the interview, as guided/prompted by the usage and exploration of the personal analytics dashboard. Interviews generally lasted between 45 minutes and 2 hours.

Online Questionnaire and Focus Group

An online questionnaire was sent to participants two weeks after the interviews in order to check whether additional thoughts and reactions would have emerged after the interviews. These questionnaires did not lead to additional insight.

The second core aspect of the study was the focus group. The idea here was that, considering that the set of participants to the study had different background, roles and views, there was value in confronting their opinions within a collective discussion. In order to most effectively collect a collective view in a very short time (90 minutes), the focus group was constructed around a specific task: The group was to generate input towards writing a business case supporting the deployment of a consumer activity data service at the Open University. This business case included two main sections: 1- benefits of consumer activity data and 2- obstacles to the deployment of consumer activity data services, and potential solutions. During the focus group discussion, participants were therefore 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 for all the participants to see as a member of the UCIAD team was adding notes to it, so that it could be corrected and validated collectively at the time the notes were made. The resulting document is available as a *Google Document*.

Findings

In this section, we summarise the main results of the study, as a set of general findings regarding the potential uses, benefits and challenges related to making available to users their consumer activity data within an organisation such as the Open University. These observations were obtained through analysing the transcripts of the recordings made during the individual interviews, together with the results of the discussions summarised in the draft business case from the focus group.

Note: In the description of the findings below, we include 'quotes' to exemplify or illustrate general notions. These quotes are taken from the transcripts of the interviews, but are not in most cases written with the exact same formulation used by participants, as they might reflect what several participants said differently, or be paraphrased to make them understandable outside the context of the full interview.

Use cases and benefits of consumer activity data

Based on analysing the answers to the interviews and the result of the focus group, we identified five main ways in which making available to users their own activity data could be beneficial to them. While these are the commonly mentioned use cases, it does not actually mean that they are shared by all users. There is also a clear feeling from the participants that, while exposing them to the possibility of obtaining their activity data leads to interesting reactions, the actual usage of such a (currently inexistent) service would really emerge from long term use, and would vary widely from one user to the other.

1. **Self reflection**: Some participants of the study saw value in simply being able to reflect on what they do, on their own usage of resources and especially in how it reflected the way they worked. These participants can often easily identify patterns in their own data, and explain them quickly (e.g., "that pick of activity on Wednesday morning is because I have my supervision meeting on Wednesday afternoon"). This notion of self-reflection on one's own activities is generally related to the one of lifelogging [2] (see in particular [3] that discusses this idea in relation to web interactions). Out of the four identified use cases however, it is the least shared, as several participants do not see much value in analysing their own behaviour.
2. **Improving the use of resources:** Very much related to the previous use case, the most commonly mentioned potential purpose of consumer activity data is to improve the use of online resources, and to make it more efficient. This includes cases where the analysis of past activities would lead to a change in behaviour with respect to interactions with online resources (e.g., "I keep looking for and coming back to the same thing, so I should bookmark it") and the realisation that the resources and services provided are not used to the best of their capacity (e.g., "I can see how much time I spend on forums, and that I could do more with them"). At a more concrete level, it is mentioned by several participants that a tool like the one provided for the study could work as an automatic bookmark system, i.e., as a way to "find again" resources that they used in the past, based on various clues (time, place, browser, etc.)
3. **Tracing anomalies:** In a way somehow more specific than the one above, a commonly mentioned use of activity data is to trace back and find evidence or information related to some kind of anomalies in the user's activities. While the two previous use cases concerned continuous usage, this relates to

punctual inspection guided by a specific goal/query or, even if it is not used at all, to the benefit of knowing that it is available to be used if any kind of situation would require it (i.e. “it is good to have it, just in case”). In the limited setting considered in the study (restricted to Open University resources), the types of anomalies to be traced are not very clear however. Participants mention scenarios such as being able to check that an activity was properly realised, or to analyse a situation that might have led to a privacy-related issue, which would be more relevant in a more general setting.

4. **Ensuring transparency:** This use case is somehow different from the three others, as it does not really relate directly to the use of online resources or even activity data, but generally to the relationship with the organisation collecting these data. Indeed, while several participants did not see the collection of activity data as being in anyway worrying (e.g. “I trust IT to do the right things, and I’m aware they are collecting this data”), it is often mentioned that it is simply “good to know what they know”. In relation to the previous use case, we are considering here the value of consumer activity data as fulfilling the user’s curiosity about what can be derived from their interactions, and as being reassuring with respect to the potential use of the data (e.g., “there is nothing there which is really an issue”).
5. **Supporting collaboration:** While there is a clear divide amongst the participants on whether or not sharing of consumer activity data, even by the users themselves, should be allowed, the value of showing and comparing analyses of activities within a group appears very clearly in specific use cases (e.g., “as a student, you could send it to your tutor for him/her to provide feedback, possibly based on comparing with others” or “you could share with a new member of the team, to show them how to do something”). Of course (as discussed in more details in later sections of this report), such usages raise a number of additional complications, related to access control over consumer activity data.

Besides the direct usage of consumer activity data by users, other general benefits are mentioned, especially to the organisation. These include in particular improving the reputation of the organisation in terms of transparency, and the trusted relationship between the organisation and its users. Related to the issues of data protection discussed below, it is also foreseen that giving to users their own activity data could produce a shift in the responsibilities in handling these data, from the organisation to the individual users, somehow simplifying the position of the organisation with respect to their data protection policies.

Requirements for delivering consumer activity data

Until now, we have been talking about the delivery of consumer activity data in very general terms, i.e., as giving it back to the user. However, of course, there can be many different ways to implement such a delivery. Participants were given access to an interface (as described earlier in this report) that provided an example of the way such data could be presented. While they were not able to comment on the technical aspects of delivering such data in this form or another, the use of the tool and the prospective idea that such data could be obtained by them in a more continuous way led to identifying specific requirements for such delivery to be effective, in four main topics:

1. **Tool support:** The most commonly mentioned requirement is that the data should come with appropriate tools to browse and query them (similarly to what was provided for the purpose of the study, i.e. "what you have here is pretty good"). Generally, the most commonly requested feature is the ability to obtain a quick overview of the data, showing general trends in the activities (therefore supporting the first two use cases identified, and well as, to a certain extent, the fourth one). In relation to the third use case, it was mentioned as important to have the ability to filter and/or query the data for specific aspects, in order to retrieve information related to the corresponding situations. It is worth mentioning that these two features were the ones supported by the UCIAD personal analytics dashboard. When explicitly asked, participants rarely judged the ability to import or integrate other sources of activity data (e.g., from social networking systems, or, if available from other organizations) as especially relevant. Similarly, while clearly supporting the fifth identified use case, features supporting users in sharing (parts) of their activity data were in general not seen as desirable. As most participants were not of a technical background, they were mostly indifferent to the idea of obtaining the data in a format that they could process themselves (e.g., XML, RDF, etc.)
2. **Completeness and correctness:** One of the most common criticisms of the data used as the basis for the study was related to its completeness. Indeed, while logs from the core servers delivering online services and resources from the Open University were included, the data did not cover other aspect of participants' interactions with the Open University, such as activities realised on departmental websites. As a result, the value attached to the activity data was seen as dramatically decreased (e.g., "I guess it would be useful, if it included everything") as the possibility to realise, at least, use cases 1 and 2, was hampered by the lack of comprehensive data. Going a step further, several participants indicated that the usefulness of activity data was more relevant in a global context, rather than being restricted to the interactions with one particular organization (e.g., "I would like to compare the time I spend on Facebook with the time I spend on the VLE[1]"), while this actually contradicts the general lack of interest for integration features in the tool. Naturally, in order to be exploitable, it is also generally admitted that the data need to be accurate. A common example of inaccuracy appeared with the location of the user at the time of interacting with Open University systems. Indeed, the location of the user is derived from the IP address of the computer used for the interaction, which is a often a misleading information. Showing the wrong location (even if only for some parts of the data) not only reduces the usefulness of the data (making location-based analysis practically impossible), it also generally reduces trust in the system and in any possible interpretation one might derive from analysing the data.
3. **Privacy and access control:** Privacy is naturally a concern when talking about personal data. While they are often not included in studies related to personal information, activity data appeared naturally to participants as representing information that ought to be protected. Participants had however

[1] Virtual Learning Environment

very different reactions to these issues. Indeed, for many of the participants, it would be important that the organisation delivering consumer activity data was to put in place the necessary mechanisms so that these data would be handled securely, as well as strict access control policies only allowing them to access their own data (even if, generally, there was rarely any information that would cause much concern to particular users). The most common privacy related issue mentioned was the risk of misinterpretation of the activities (e.g., "someone might think that I'm looking for another job"). Other participants were more relaxed about the general issue of privacy (e.g., "there is nothing there that I would not share myself"), but still indicated that they would require having control over any potential distribution of the data to third parties. When prompted, participants admitted that there might be scenarios in which the data could be used for malevolent activities (e.g., identify theft, stalking, etc.), or even to obtain confidential information from the organisation by external parties (e.g., structure of internal sites, possible weakness in security infrastructure, etc.) These scenarios were however often judged unlikely, especially under the assumption that sharing data other than with the corresponding individual users was to be prevented.

4. **Cost effectiveness:** Handling and delivering data of course comes at a cost, including the cost of storage, transfer, maintenance and security. From the organisation's point of view, it remains to be evaluated how much of an additional cost the delivery of consumer activity data would be, considering that such data is already being collected and processed for the needs of the organisation. It is important to also consider here the cost to the user: depending on the mode of delivery, consumer activity data might need to be stored by the user, together with the tools provided to process them. The delivery format, the tools proposed and the mode of delivery would therefore have to be designed in such a way that it minimises overhead for the user (ideally, being done transparently), so that the effort required does not overcome the benefits (or, in one of the participant's words, "life's too short, so you might look at it, and then just move on to something else").

Policy level implications

Implementing consumer activity data delivery within an organisation would have one major impact in terms of the policy to put in place: it would change the status of the data. Indeed, activity data are already being collected by most organisations, in web server and web applications logs like in our study, or through other forms of analytics and monitoring systems. These data are however generally considered internal data, to be used as supporting information for maintenance (debugging, system monitoring) or to improve the services provided to users (through adding or modifying features detected as inefficient, or through the usage of "collective intelligence" techniques [4] to automatically generate personalised, relevant entry points to available resources and services).

The current status of activity data actually results in that it is not even clear whether it should count as personal data. Indeed, the *Data Protection Act* which applies to any personal (identifiable) data collected by organisation stipulate that privacy notices should be stating by who, how and why personal data is being collected, and possibly indicate how such data can be accessed and corrected. All of the participants to our

study were unaware of this type of information in relation to activity data collected at the Open University. Generally, most privacy statements from organisations other than the ones using activity data as a core asset (such as Google) remain vague in their general privacy policies regarding the definition of personal data, and it is unclear whether activity logs are considered part of somebodies personal record, even if the *Data Protection Act's definition of personal data* clearly applies here[2]. It is interesting in particular to note that, when asked, the study participants were all unsure whether they could claim the right to access the activity data (such as server logs) collected by the Open University about them, were unaware of the specific data protection and retention policies in place, and did not think that they could request for such data to be deleted or corrected. They were generally unsure whether the Open University was transferring (or had the right to transfer) such data to third parties, but expected that they should, at least, be alerted if that was the case (e.g., "I don't know, but I hope they don't"). As a way to generate reflection on the general complexity of this area (data protection and privacy), we also asked participants in relation with the previous question of transferring data to third parties, what they would think if the Open University were to use *Google Analytics* on the considered websites. The idea here was to think of the use of Google Analytics as having for side effect that activity data they (as users of Open University websites) generated was transferred to a third party (namely, Google). Most participants first did not know whether or not this situation applied (as far as we know, it does not) and, as expected, indicated that they had not thought of the issue in this way. The general reactions obtained as a result confirmed in particular that the complexity of notions related to data protection related to activity data and the general vagueness associated with them were detrimental to users, and that it was important to achieve more transparency in the collection and use of activity data, with providing access to consumer activity data being one way forward in this area (as discussed in the fourth use case above).

As the current situation with respect to activity data is already rather confusing, it appears clearly from the discussions with participants that moving towards consumer activity data would require clarifications, and possibly new arrangement, for what we refer to as "ownership of the data". Indeed, briefly stated, it will be important whenever consumer activity data services are put in place to clarify explicitly the rights associated with data for both users and the organisation. When asked about their expectation regarding data ownership arrangements, study participants came up with four different possible approaches, representing different points in a spectrum ranging from "full control to the individual" to "full control to the organisation". We describe these approaches together with some considerations regarding their feasibility and rational below, starting with the extreme cases[3]:

1. **The user owns the data:** This approach (which represents the most extreme change with respect to the current situation) is one where the data are being

[2] It is worth mentioning that server logs as considered in our study are currently stored securely and only kept in a non-aggregated form for 7 days at the Open University.

[3] It is worth mentioning that the ownership of data is in itself a rather fuzzy notion. Copyrights for example do not apply to data, as they are not creative work. Here, we assume that being the owner of data relates to "*database rights*", and so to generally being able to allow and prevent access to data by external parties.

transferred to the individual with full rights on them. The rational behind this approach is that, since activity data represent personal information generated as a result of the user's activities, it is only natural for the individual to be the one in control of it. One obvious objection to this approach is that it means that the organisation would not have by default the right to exploit activity data (or even access them), without explicit agreement (i.e., a license) from the user. It would naturally make more complex any genuine use of activity data (debugging, analytics) and would make the collection of activity data irrelevant for organisations. Other objections includes that the organisation should naturally be the owner of the data, since they are generated out of processes and resources owned by the organisation (which is especially relevant if the user is an employee of the organisation, as often the case of our participants).

2. **The organisation owns the data:** As far as our legal knowledge goes, this is the current situation: The organisation that is generating the data detains the rights on these data and can exploit them as its own asset. These rights are however generally restricted by the Data Protection Act, meaning that there are only certain possible usages and exploitation channels that can be applied (at least without consent from the user). In this case, delivering consumer activity data would effectively mean issuing a license to the user for him/her to have access to users to their own data. An advantage of this approach (from the point of view of the organisation) is that the license granted to the user can include restrictions in the way the data might be used, therefore protecting the organisation's asset as well as any potentially confidential information that might be gathered by external parties from aggregating activity data. It is worth mentioning that most of the study participants indicated that they would find it appropriate for the organisation to impose strong limitations to the use of the data (e.g., that they could not be shared and could only be accessed through the tools provided by the organisation). An obvious objection is that such restriction would also impose strong limitations to the potential usefulness of the data.
3. **Data co-ownership:** An intermediary approach would be to consider both the organisation and the user as owners of consumer activity data. This means in particular that both would have full access to the data, but should be prevented from using them in ways that could be detrimental to the other party. While the legal implementation of such an arrangement could be rather complex, it might be achievable considering that the organisation would still have to comply with the Data Protection Act regarding the co-owned activity data, and users might be bound to confidentiality agreements preventing them from widely distributing these data, since they might represent potential threat to the business or the security of the organisation.
4. **Transferring ownership instead of deleting:** The Data Protection Act indicates that personal data might not be kept by an organisation for longer than necessary for the intended purpose. As already mentioned, as activity data is mostly used by the Open University for the purpose of system maintenance and for aggregation into analytics, they are only kept for 7 consecutive days. One possible approach would therefore be to transfer the data, together with all associated rights, to the user after this retention period. The rational here is that, since the data would normally be deleted, they do not represent anymore an asset for the organisation, while they might still be of

use to the user. From a practical perspective, this would also represent a convenient arrangement, as users could simply ‘opt-in’ for their consumer activity data to be sent to them at the time it is deleted by the organisation. Of course, this would still imply that the organisation would lose control over the data once transferred to the user, but similar “non detrimental use/confidentiality” clauses could be attached to the delivery of consumer activity data.

Conclusions

The current trend towards consumer data and the general raise in awareness by Web users of the importance of appropriately managing their personal data is leading to more and more demand for services giving back data collected by online organisations to the individual users of these organisations. The ways in which such services will eventually emerge is still unclear however. Indeed, foreseeing the particular usage individuals will make of such data is an impossible task as long as such services are not widely deployed. It is also expected for such a trend to generate immense challenges, not only at the technological level, as it will require for organisations to re-define the way personal data is handled and delivered.

In this report, we focused on the particular area of activity data, i.e., data generated out of the interactions between individuals and online services or websites. Based on a prototype “consumer activity data” system, we realised a user study to identify the main potential usages of activity data by the users of these services and websites, the requirements for the delivery of such consumer activity data, as well as the policy level implications that the deployment of consumer activity data services would have on organisations. Through this study, we describe both the benefits that are expected from the implementation of such services, as well as the major obstacles that are hampering their realisation. More precisely, we mainly conclude that 1- web users are expecting such services to appear, and the usages for them to emerge our of their availability; 2- The already identified use cases and organisational benefits appear to justify the cost of opening up consumer activity data; 3- Such services will generate profound changes in the way data is handled and controlled by organisations and; 4- Additional work is required to achieve an appropriate development of both the technological and the organisational infrastructures around activity data as personal information, in anticipation of the appearance of consumer activity data services.

References

1. d'Aquin, M., Elahi, S. and Motta, E. (2011) Semantic Technologies to Support the User-Centric Analysis of Activity Data, Workshop: Social Data on the Web Workshop, SDoW 2011 at ISWC 2011
2. O'Hara, K., Tuffield, M. and Shadbolt, N. (2009) Lifelogging: Privacy and Empowerment with Memories for Life. Identity in the Information Society, 1, (2)
3. d'Aquin, M., Elahi, S. and Motta, E. (2010) Personal Monitoring of Web Information Exchange: Towards Web Lifelogging, Poster at Web Science 2010 Proceedings of the WebSci10: Extending the Frontiers of Society On-Line
4. *Segaran*, T. (2007) Programming Collective Intelligence, O'Reilly Media, ISBN: 978-0-596-52932-1