Towards a Mobile Social Data Commons

Giles Greenway*, Leonard Mack[†], Tobias Blanke*, Mark Cote* and Tom Heath[†]
*Department of Digital Humanities, King's College London, United Kingdom
[†]Open Data Institute, London, United Kingdom
Email: tobias.blanke@kcl.ac.uk, giles.greenway@kcl.ac.uk, mark.cote@kcl.ac.uk

***Abstract*—This paper discusses how born-digital cultural material can be opened up for research. We focus in particular on the grey area between private mobile phone data and its publication and use for research and beyond. We report on the results of the 'Empowering Data Citizens' (EDC) project, which is a collaboration between King's College London and the Open Data Institute. The work builds on the project Our Data Ourselves (http://big-social-data.net/), which studies the content we generate on our mobile devices, what we call big social data (BSD), and explores the possibilities of its ethical storage.**

I. Introduction

Our project addresses a basic research question: How do we transform BSD into open data, and in turn, empower the end users of mobile devices and cultivate new data communities? There are basic contradictions here that necessitate our cultural-technological approach. On the one hand, (meta)data is private data, digital traces identifying who, what, where and when. This is data already deeply embedded in digital enterprises and the security state as a source of both economic value and surveillance. Yet, it remains largely out of the hands of the everyday use of those who produce it. On the other hand, metadata is more than just a source of economic value or surveillance; it reveals a surprising breadth and depth of cultural activities. In identifying the who, what, where and when, these digital traces offer innovative approaches to the core of arts and humanities research.

Once understood as mere machine-to-machine data because it is composed of communication between a device and a cell tower, GPS satellite, application, and so on, standard mobile phone metadata can provide nuanced insight into cultural life, as NSA/GCHQ revelations show. We identify both an emerging cultural phenomenon and a technological challenge. The cultural phenomenon is the effortless and ubiquitous production of metadata that is increasingly able to provide intimate details of social and cultural life. The technological challenge is in transforming that machine-to-machine data into interoperable open data, but in a manner that safeguards privacy and facilitates research.

Consequently, our aims and objectives are to research the implications of opening cultural data, investigating new data cultures and develop tools and apps for their cultivation.

II. Background

The work we report upon has contributed to our existing work twofold. First, we have developed an open linked data framework to effectively embed anonymised born digital cultural data. Our technological research has developed proof-of-concept demonstrators that investigate the use of advanced anonymisation technologies for publishing cultural data. This approach will cultivate open data cultures, for example, by presenting the potential surplus from integrating it with other linked data resources such as the concept ecosystem of DBpedia [1].

The main vector of our research is in approaching born-digital cultural content via the model of open data. Open data refers to data available for anyone to use for any purpose and free of cost. Open data should be in formats that are interoperable, that is, it can be linked, and thus easily shared, in a standard and structured format for easy reuse. The key deliverable of our project is the cultivation of an ethical environment of openness for this kind of important born-digital content for cultural analysis.

To this end, we first need to address what seems to be mainly a technical problem, namely the necessary anonymisation to safeguard privacy. Yet, there is a cultural imperative if it is to function as open data which necessitates it being rendered functional and interoperable with other open datasets. Finally, there is a need for it to be open to researchers for critical inquiry. This requires us to develop an ethical and dynamic environment of information sharing via our techno-cultural interdisciplinary approach. In turn, this will facilitate on-going and innovative research into the datafication of all human conditions. While there has been extensive research into anonymisation techniques and open data in a range of other academic disciplines from health-care to social sciences, this research closes this gap for arts and humanities research on contemporary born-digital material. There is very little research to our knowledge in the field of anynomisation of cultural data. Where there is existing work, it mainly relates to heritage data sets, rather than born-digital material, or related to social science repositories.

In summary, then, our aims and objectives are as follows:

1) Understand the conditions of possibility of opening BSD and developing a road map for future research on open BSD
2) Experiment with technologies and methodologies to facilitate research on this new born-digital material and explore open data cultures
3) Theorise the changing conditions of open data culture and thus empower data citizens

III. The Original MobileMiner Application

The Our Data Ourselves project has engaged young coders in the co-research and development of innovative analytical tools and methods to expand this new area of cultural research. The EDC project will develop proof-of-concept demonstrators that investigate the use of advanced anonymisation technologies for publishing cultural data. Anonymisation, however, can only be the first step. The second one has to be the investigation into what cultural research can still do with such an anonymised resource. If the fundamental humanistic question is why something happened, this why enquiry will generally based on the who, where and/or when. These are, however, exactly the kind of attribute that computerized and manual anonymisation techniques target. We need to exactly map out what is left for the scholarly discourse if we cannot make these kinds of links. Therefore, we will develop a framework for open linked data to effectively embed anonymised born digital cultural material.

Previously, we have discussed the development of our "MobileMiner" application for Android smartphones [2]. This was in response to our question as to whether the data trails produced by other phone applications could be collected into a "social data commons" while still respecting users' privacy. The application was installed on smartphones given to twenty young coders from Young Rewired State [3], who we regarded as collaborators. Even if full access to the network traffic of a device were desirable, it would require root or administrative privileges. This would be a significant barrier to any future large-scale adoption of the application, and encouraging users to root their devices could leave them vulnerable to malware. Software that exploits vulnerabilities to grant users root access has been shown to be similar to some examples of malware. It is plausible that some apps that claim to grant root access altruistically have malicious payloads [4].

We were able to track when other apps opened or closed network sockets on Android devices by polling the `/proc` directory of the underlying Linux filesystem. This also provided the TCP port of the socket, mostly of interest in determing how frequently secure HTTPS traffic on port 443 was used. The Android API provides a log of network usage on a per-app basis, and this was also polled. Finally, the application requests that the user authorizes it as an accessibility service, which gives it access to notifications sent by others. A legitimate accessibility service might render notifications in large print or read them out. Only the times of the notifications and the applications that sent them were logged, the text they contained could include parts of emails or other private communications, and were deemed too invasive to collect. The interpretation of this data is highly dependent on the application that generated it. As we have seen previously with mobile games, [2] socket usage can be a reasonable proxy for a user's interaction with an application, but some make contact with servers so frequently that they tell us nothing about the user.

Although these methods track the activity of mobile applications, they provide no information about what is transmitted or received. It is possible to reason about this given the permissions an application requests. Information about location via GPS and mobile and wireless networks is of particular interest, but no way to determine when and where this has been accessed has been found for a standard non-rooted device. To illustrate what location data allows applications to learn about the lives of users, MobileMiner also logs the IDs of mobile cell towers and wireless networks that devices connect to. The networks that are merely visible but with no connection are not logged, neither are GPS or Google's location APIs used; they were deemed too invasive. The cell tower IDs were converted into approximate locations using the OpenCellID database [5]. This often provides a reasonable degree of obfuscation, such that only a user's very general habits are disclosed, but in some cases specific places of work or study could be identified. The SSIDs of wireless networks often contained the names of institutions that provided them, revealing information about users more explicitly.

Users are able to start and stop the recording of data at any time. In the first iteration of the application, new data is uploaded to a CKAN server [6] via a custom plug-in every ten minutes while the user's device has a wireless internet connection. The application gives the option of copying its internal SQLite database to an area of the device's internal storage accessible when it is mounted as a mass-storage device so that users can access the data they generate. So far, access to the CKAN database has been granted very selectively on an individual basis. We have held a hack-day for users of the application [7] where a virtual machine image of the CKAN server and its data was shared as a Docker container. Data was manipulated and displayed by providing a browser-based Ipython notebook with the MatPlotLib library installed within the virtual machine.

IV. Enriching MobileMiner with OpenPDS

The MobileMiner application is conservative in the data it collects. This is in sharp contrast to the Funf Open Sensing Framework for Android, developed in part by the MIT Media Lab [8]. This provides a set of "probes" that captures a wide range of data, and the means for its short-term storage and eventual dissemination to an external server. Probes include periodic capture of the IDs of nearby Bluetooth devices, full GPS location, and mobile web-browser history, far beyond the remit of MobileMiner to examine the behavior of applications rather than that of users. Rather than collect only data with a low risk of identifying the user, the Open Personal Data Store (OpenPDS) architecture was proposed by the MIT Human Dynamics group [9] to enable the querying richer data without betraying users' privacy.

We have adapted the MobileMiner app to provide the option of gathering data via the Funf framework and transmitting it to an OpenPDS instance. Custom Funf probes that call the pre-existing code that polls network sockets and traffic usage API were implemented. The decision that MobileMiner should only log active connections to cell towers and wireless networks, rather than periodically scan for available ones requires the development of other custom Funf probes, rather than the use of existing ones that capture this data. An overview of the flow of data is given in figure 1. MobileMiner uses a custom Funf pipeline configured to transmit data to a Personal Data Store (PDS) using the OpenPDSClient Android library. Access to a PDS is mediated by a registry server, we have made Docker images available so that both pieces of software can be deployed for experimentation and testing very easily. The intention is that third parties authorized by the owner of the PDS query the data by submitting questions to it such that only summarized answers are returned, rather than the raw data. This is achieved by submitting third-party Python code to the PDS which is executed using the Celery task scheduler. The potential for submitting malicious code that betrays the

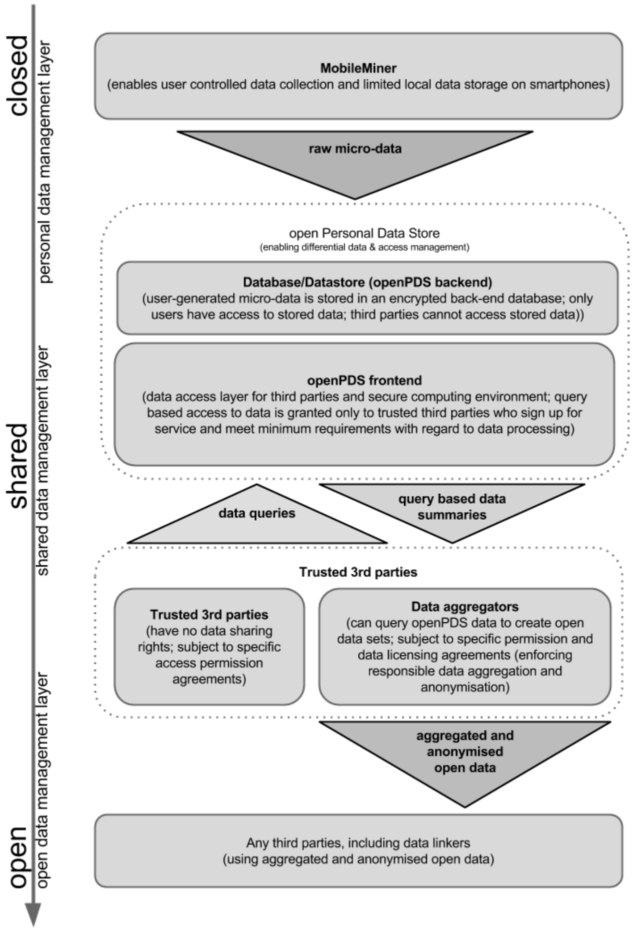
closed
personal data management layer
MobileMiner
(enables user controlled data collection and limited local data storage on smartphones)
raw micro-data
open Personal Data Store
(enabling differential data & access management)
Database/Datastore (openPDS backend)
(user-generated micro-data is stored in an encrypted back-end database; only users have access to stored data; third parties cannot access stored data))
openPDS frontend
(data access layer for third parties and secure computing environment; query based access to data is granted only to trusted third parties who sign up for service and meet minimum requirements with regard to data processing)
shared
shared data management layer
data queries
query based data summaries
Trusted 3rd parties
Trusted 3rd parties
(have no data sharing rights; subject to specific access permission agreements)
Data aggregators
(can query openPDS data to create open data sets; subject to specific permission and data licensing agreements (enforcing responsible data aggregation and anonymisation)
aggregated and anonymised open data
open
open data management layer
Any third parties, including data linkers
(using aggregated and anonymised open data)

Fig. 1. Data workflow for MobileMiner and OpenPDS.

privacy principles of OpenPDS or otherwise undermines the security of the machine on which it is hosted has yet to be fully explored. A restriction of the modules from the Python standard library that such questions are allowed to import would be a good starting-point. The Python runtime environment of Google's "App Engine" cloud computing service provides a modified version of the "os" module that restricts access to the underlying operating system. Python's abilitity to manipulate the byte-code to which it compiles could also be exploited [10]. At present, the registry server software specifies a hard-coded location for the PDS server in its configuration file. Clearly, OpenPDS is some way from a safe mass-deployment by consumers on, for instance, smart TVs, home media servers or popular low-power devices such as the Raspberry Pi, and we consider its use as experimental.

V. Privacy Challenges

The project is on-going and will conclude in October 2015. The core research interest of EDC is to explore the boundary between personal and open data. Subsequently, the project aims to identify workable solutions which enable individuals to execute greater agency over their data when handling, trading, and releasing it. Whether individuals will adopt such tools will likely depend on effective and credible means for privacy protection. Preparing such data for a publication through a structured end-to-end process which enables data citizens to manage the data through its entire life cycle and protects privacy is a particularly complex issue.

In the EDC workflow model, data flows from individuals via the MobileMiner app on their smartphones to a database, which third parties query-based access to the data. From there, data aggregators can also query data to create open datasets. The privacy challenges of this model are, at least, twofold. Firstly, from an individuals perspective, it makes it much harder to define acceptable privacy boundaries. As Helen Nissenbaum [11] argues, individuals chose what they want to reveal dependent on context. Hence, publishing micro-data, to an undefined audience with undefined capabilities and for undefined purposes requires a robust and credible privacy protection framework handled through human or technical intermediaries. Secondly, even for such a governance approach, the unknown data users and their unknown technical capabilities are a significant problem. The core issue here is that with openly shared data, the data users cannot be known ex ante and cannot be controlled. Consequently, common ethical and normative standards for privacy protection are hard to define and probably not enforceable. Additionally, it is unclear which technical capabilities the audience of data users has; which data resources they can use to produce more privacy-harming linked data with a higher degree of unicity or which capabilities they have to reverse-engineer applied anonymisation techniques. Like many other smartphone apps, the MobileMiner app collects a number of microdata, such as location data, data on the use of mobile phones and specific apps installed. Hence, to what extent the specific data collected by the app can be used to re-store privacy invasive information both only from the data itself and in combination with outside information is a core question for EDC.

Enabling wider sharing of data first requires an assessment of how the data collected by MobileMiner might cause privacy risks. The core issue here is that data generated by the MobileMiner app, much like any other data produced by smartphone apps, potentially represents a set of micro data with a high dimensionality and sparsity. Hence, preparing such data for a publication is a difficult task, if privacy is to be protected. Therefore a detailed privacy risk assessment was conducted on all data which is currently collected by MobileMiner. The objectives were to assess the entire body of data collected by MobileMiner, evaluate whether and to what extent the data impacts privacy, and suggest solutions for risk mitigation where appropriate.

Overall, it appears that the data collected by MobileMiner is mostly non-sensitive and broadly relatively unproblematic. The most novel data classes collected, network socket and traffic usage, are particularly innocuous, with little chance of betraying the identity of users. MobileMiner mostly follows a conservative data collection approach. This limits the necessity for additional actions to reduce privacy impacts. We should however add, that this assessment is made under two assumptions: The first one is that users are sufficiently informed when they decide about any transfer of MobileMiner. As described in the end-to-end data workflow, this should be achieved by a thorough provision of crucial information to ensure informed consent, including clearly stated terms and conditions, privacy notices, and contributor agreements (in cases where data is to be used by data aggregators that create open datasets). The second underlying assumption is that, as it is currently the case, no demographic data on the users of MobileMiner or openPDS is collected and/or shared with third parties. Sharing data such as age or gender with third parties might on the one hand greatly increase the utility of the data provided. On the other hand, it would however also increase the ease of reidentifying individual users.

With regard to individual data classes, two aspects need to be highlighted: First, MobileMiner generally collects time-stamped location data, which could be used to distill unique movement patterns [12]. However, location data collected by MobileMiner is rather sparse. The app does not access GPS or Googles location APIs, it only records data on the cell-towers to which the phone was connected at a given time. This undermines any efforts on cell-tower triangulation. Instead, it only allows to locate individuals with a proximity of, at best, a few hundred meters in urban spaces with a high cell-tower density. In rural areas, this might easily be expanded to a space of several hundred meters. Accordingly, MobileMiner only allows to create very rough movement patterns. Mobile phones can connect to a series of towers even while stationary, clustering such a time-series of locations may yield a slight increase in precision. We therefore assess that the location data by MobileMiner data only creates low privacy risks. Re-identifying individual users without substantial amounts of external information about them appears to be rather difficult under these circumstances. Rather, in line with EDCs general research interest, this data could be used to assess aggregate movements of user groups. Second, the most invasive data collected by MobileMiner is the smartphones wireless network data. Here, the plain text of wireless networks which the smartphone connects to could greatly increase privacy impacts. If a phones regularly logs into a wireless network named “TestSchool”, it is very likely that the owner of the smartphone is either a pupil, teacher, or other member of school staff. This again might allow an external attacker to identify additional links which eventually help to identify a real person behind the data. Certainly, the privacy impact of this data being accessible largely depends on an external factor, i.e. the name of a wireless network. However, to reduce potential privacy risks, we should consider removing this class of data entirely from the dataset or replacing wireless network names with a unique identifier.

VI. Conclusion and Future Work

The initial release of MobileMiner collected a fixed set of data classes, which were only made available to the user if they were also transmitted to us. So far, the only users of the application have been individuals closely involved with the project known to us personally. From the client side, all that is needed for a large scale release is fine-grained control over which data classes are transmitted, the obfuscation of wireless network data and the embedding of clear terms and conditions within the application. We shall demonstrate MobileMiner interacting with an OpenPDS server via the Funf framework and update our risk assessment in the light of this. For large-scale adoption, OpenPDS will need to demonstrate in principal that it can resist attempts at de-anonymisation with realistic computational resources if queried as intended. It will also need to be shown to be resistant to malicious requests in practice.

References

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, “Dbpedia-a crystallization point for the web of data,” *Web Semantics: science, services and agents on the world wide web*, vol. 7, no. 3, pp. 154–165, 2009.

[2] T. Blanke, G. Greenway, J. Pybus, and M. Cote, “Mining mobile youth cultures,” in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 14–17.

[3] “Young rewired state website,” 2015, [Online; accessed 27-August-2015]. [Online]. Available: http://www.yrs.io

[4] W. Georgia, “Bypassing the android permission model,” 2012, [Online; accessed 27-August-2015]. [Online]. Available: https://www.youtube.com/watch?v=n0DlYumayGc

[5] “Opencellid website,” 2015, [Online; accessed 27-August-2015]. [Online]. Available: http://opencellid.org

[6] T. O. K. Foundation, “Ckan — the open source data portal software,” 2015, [Online; accessed 26-August-2015]. [Online]. Available: http://ckan.org

[7] J. Pybus, “A long overdue update on the success of our second hackathon!” 2015, [Online; accessed 26-August-2015]. [Online]. Available: http://big-social-data.net/2015/01/10/a-long-overdue-updates-on-the-success-of-our-second-hackathon/

[8] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, “Social fmri: Investigating and shaping social mechanisms in the real world,” *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643 – 659, 2011, the Ninth Annual {IEEE} International Conference on Pervasive Computing and Communications (PerCom 2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1574119211001246

[9] Y.-A. de Montjoye, E. Shmueli, S. S. Wang, and A. S. Pentland, “openpds: Protecting the privacy of metadata through safeanswers,” *PLoS ONE*, vol. 9, no. 7, p. e98790, 07 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0098790

[10] J. Geralnik, “Pytroj. a tool for infected .pyc files with arbitrary code that spreads out to infect all other .pyc files,” 2012, [Online; accessed 26-August-2015]. [Online]. Available: https://github.com/jgeralnik/Pytroj

[11] H. Nissenbaum, “Privacy as contextual integrity,” *Washington law review*, vol. 79, no. 1, 2004.

[12] Y.-A. de Montjoye, M. Hidalgo, Csar A. Verleysen, and V. D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific Reports*, vol. 3, p. 1. [Online]. Available: http://www.nature.com/articles/srep01376