Designing Data Retrieval App to Study Facebook User Participation

Anja Bechmann
Aarhus University
DK-Helsingforsgade 14, 8200
Aarhus N
anjabechmann@gmail.com

Peter Vahlstrup
Aarhus University
DK-Helsingforsgade 14, 8200
Aarhus N
imvpbv@gmail.com

CHI'13, April 27 – May 2, 2013, Paris, France.

Abstract
The purpose of this paper is to shed light on the opportunities and challenges of using Facebook's open API to do qualitative studies of user interaction and communication that happens behind the login. Through design of a Facebook data retrieval app the paper contributes to the field of HCI by discussing optimization and methodological obstacles that arise in qualitative studies of user interaction that happens behind authentication (Conditions of use, legal and ethical considerations, and interface differences between the actual user interface and the data retrieved through the open API).

Author Keywords
Open API; Facebook; Qualitative user studies; data retrieval, data statistics, application design

ACM Classification Keywords
H.5.m: Miscellaneous.

Introduction
Statistics show that increasingly much of the time spent on the internet happens on or through social networking sites where user engagement unfolds in password protected environments. To make qualitative studies of user participation in these environments we

need to make these interfaces and the communication and interaction going on here objects of study.

On Facebook and connected partner services such as Instagram and Spotify methodological tools such as web crawling [11,5] is not an option because it does not allow us to crawl pages that require user authentication. Raw data logfiles are the ideal objects of study in these enviroments, but they are only accessible to qualitative social media researchers if they are employed by the social media company or an extended collaboration is established with the social media company in question [18] and this has made it nearly impossible to analyze what is going on inside Facebook, in terms of data and information patterns from an outside perspective. Can we access Facebook user interaction and communication otherwise? And how do we make sure that qualitative researchers without programming skills can use the methodological tool?

The aim of this paper is to shed light on the possibilities and challenges of using the open API of Facebook to make qualitative studies of user interaction and communication happening behind login. The paper contributes to the field of HCI by discussing how we can optimize qualitative studies of user interaction that happens in authentication settings and which methodological obstacles arise in terms of company's terms of use, legal and ethical considerations, multiuser app interface, and interface discrepancies between the actual user interface and the data retrieved through the open API of Facebook.

The paper will outline and discuss existing methodological tools for qualitative researchers to retrieve and study Facebook data, followed by a description of the multiuser app we designed to retrieve data from Facebook through the open API. From this app design case we will discuss the possibilities and obstacles of studying Facebook user interaction through open APIs and app for multiple researchers.

Related work

In this section we have chosen to outline related work by focusing on: a) similar retrieval tools b) the use of open APIs as data retrieval tool and c) the use of classical qualitative methodological approaches on Facebook data retrieval and analysis.

Existing social media data retrieval tools

Some large-scale analytical systems serve as platforms for accessing data on Facebook, but they only crawl publicly available data, through public profiles e.g. SocialMention (http://socialmention.com/) and Radian6 (http://www.radian6.com/). The purpose of these services is for companies to scan the social media networks for mentionings. These systems scan across social media sites, but on Facebook they only crawl data from public profiles. This, despite the fact that Facebook allows third party companies and researchers to access user data with the permission of the user, if the third party company/researcher complies with Facebook's terms of use for developers.

NVivo is a qualitative data analysis system that among others facilitates retrieval and analysis of Facebook data (http://www.qsrinternational.com/). The software give access to semi-private data, but through the browser plugin NCapture analysts can only access public profiles and groups or posts from users they are

friends with and groups they are part of. Furthermore NVivo cannot import the newsfeed.

ArchivedBook is an app using the Graph API for data retrieval, but ArchivedBook is not a tool for researchers. It only shows the end-user's personal data. It is not possible nor the purpose to retrieve data from other users. As a researcher you need a system that ask for permission to retrieve data that are not publicly available from a sample of end-users.

Open APIs as Data Collection Tools

In recent years there has been a focus on medium-sized and "Big Data" structures in cloud services, and how web crawlers, databases, and graphical presentation tools can help surmount the obstacles to documenting user interfaces, making data accessible and searchable for data mining, for example, the use of open Application Programming Interfaces (APIs) to extract data from social network services.

Quite a few researchers studying open/public social networks have used open APIs on a smaller or larger scale to extract and crawl data. This is particularly the case with Twitter studies, where using open APIs has been the standard approach for extracting data to various external software, such as IssueCrawler and other graphical network tools [4,20]. Public data has also been accessed for YouTube studies, to sample survey users on the basis of data from the Google Data API [8].

However, few, if any, studies have used the Facebook's Graph API to crawl private data with user consent, or in other words, to act as a third party, accessing the users' private information, which can inform researchers of actual user behavior, making mixed methods study possible, triangulating with ethnographic methods (observation, interviewing, etc). Other researchers have used open APIs in specific studies on other social networking sites. Korn et al. [17] use check-in data in their study of location based data patterns on Foursquare. This method enabled them to automatically sent text messages shortly after the participants checked in, and they also triangulated methods, interviewing participants afterward, to obtain rich, in-depth accounts of user experiences (rich deferred contextual interviews) [7,3].

A common denominator for these open API studies is that they are customized for specific purposes and often made by assisting computer scientists. In contrast social scientist and humanist qualitative researchers need a generic multiuser system that does not require collaboration with computer scientists in order to retrieve data from the Graph API and design the database structure. In return such system will not encompass customized needs.

Classical Qualitative Methodological Approaches

Traditionally, methodologically approaches for studying virtual communication have involved various kinds of ethnographically inspired approaches, such as classic participatory observation [12]. Classic physical participatory observation follows the user in real time, but the method is time-consuming, and very intrusive. The activity on social networks happens both during the day and night, during work and leisure hours, making comprehensive observation difficult. The researcher has to discuss the viability of observations, as the presence of the researcher redefines the situational context in which the use of Facebook takes place [3].

Another method used is technology supported diary studies [3,6], and observations using other technically supported tools, such as screen capture client software eg. Jing (http://www.techsmith.com). Facebook use happens across devices, and tracking software can be installed on the different devices and combined with a unique user identifier, so that it can track across devices, as is the case with services from Apple, for example. However, the data generated by this tracking software is in the form of video recordings of screen activity, which is weak in terms of rendering them searchable within a database structure.

Online virtual ethnographers [15,21,13,1] conduct various kinds of participatory observation studies by participating in the online community, for example, connecting to/"friending" research participants, commenting on updates and discussions, and asking for elaborations on different actions in the community. As a method, virtual ethnography builds on the fundamental ethnographic idea of becoming involved in the community studied, but has been critized for being to big a part of the community studied that researchers lack detachment necessary to see and follow up on interesting aspects of the patterns of use and communication taking place. Furthermore, the studies cannot describe how personal user data influence the user's experience. For example, on Facebook, the virtual ethnographer must rely on the Facebook algorithm showing all participant activity, and the researcher cannot see the participants' news feed, and therefore cannot see what s/he is exposed to. To document the activity, the virtual ethnographer must manually copy/paste the content from Facebook, in order to analyze it further, and cannot access the participant's news feed. This in return is one of the main issues of designing a open API data retrieval tool, to make data searchable and sortable for analysis.

Case study: Designing Digital Footprints as a data retrieval tool

The initial aim of Digital Footprints is to make it easier for researchers to collect and analyze Facebook data (private, as well as public) from a selected number of participants. Compared to existing methods we needed a tool that among others could be more unobtrusive, make data searchable and sortable, access historical Facebook data, and help secure informed consent.

\Requirements & functionality: Method:	Programming skills	Informed consent	Time consumption for researchers	Time consumption for participants	Searchable and sortable data	Accessing historical Facebook data
Physical Participatory Observation	Low	Not facilitated	High	High	No	Yes
Technology supported diary studies	Low	May be facilitated	High	High	No	No
Screen capture software	Low	Not facilitated	Low	Low	No	No
Friending participants	Low	Not facilitated	High	Low	No	Yes
Customize a new solution	High	Not facilitated	Low	Low	Yes	Yes
Digital Footprints	Low	Facilitated	Low	Low	Yes	Yes

Tabel 1. Digital Footprints compared to existing social media methods.

The application is designed to optimize small-scale, in-depth, qualitative user studies with the intent to study a small sample or panel of users. Inspired by the seven steps in qualitative studies [21], the system is designed to address the following steps: 1. To make a new Facebook research project/panel.2. To automatically guide and collect the legal information needed from the researcher and the user 3. Inviting participants to take

part in a project/panel 4. Retrieve the Facebook data needed from the participants 5. Facilitating different analytical functionalities such as data statistics, search and sorting possibilities.

New research project
Through an authentication process, researchers conduct their own research project on the web interface of Digital Footprints, accessing our servers directly. The identification process exists to identify the researchers as such, for example. Thereafter, the researcher will be able to administer his/her research projects. The administrative researcher can e.g. change the project status, update data and delete data etc.. Furthermore, for privacy protection and as a security measure, it is important for us as developers to be able to approve or deny a research project.

Legal information
In studies using open APIs [8], legal considerations on privacy have not been explicitly addressed and discussed as a part of the study and the system design. This is extremely important to us when using open APIs. We have consulted privacy lawyers, to build in the legal aspects of working with private, potentially sensitive user data in our system. Hence, an important part of the system is providing users with accurate information (e.g., compliance with EU law) about the research project in which they are participating. Even though the system is intended for international use, at this time the legal guidelines are limited to EU law [9,10,24].

For this reason, it is mandatory for researchers to furnish information on the duration of the project, the purpose of the project, the data needed for the project, and how and under what circumstances they will analyze the data. Furthermore, should data not be deleted at the end of the project period (this will be done automatically by the system), researchers must verify that any archiving is done according to the law, in other words, transferred to a state library [9,10]. According to the law, a user must be able to withdraw from a research project. Therefore, the system has a built-in function that allows participants to select "delete profile from research project". An e-mail is sent to the researcher, so that s/he can contact the user, in order to determine why the deletion took place before the data was deleted from the system. This is done because small-scale, in-depth studies are more sensitive to user withdrawal than are large-scale studies.

Inviting participants
The researcher invites participants to take part in a research project or panel by providing information required by law, but implemented in the system as standard information provided, in order to start inviting participants. The user consent process is threefold. First the researcher sends out email or recruits in person (depending on methods and panel size), informing participants of the research project (purpose, aim etc) asking potential participants to visit a URL unique to the system. Second, this URL directs potential participants to a web interface (at this time, using a browser interface) where they receive more information on the research project along with privacy policies and are asked to log in with their Facebook sign-in password. The participants then give their permission to draw data from their profiles – to the research project.

Collecting data
When a user permits Digital Footprints to collect the data specified in the invitation, the data is transferred through Facebook's open API to the server as JavaScript Object Notation (JSON).

During the initial test phase, the data collected from Facebook was private data normally only accessible in closed networks, such as user data (name, e-mail address, location, religion, political views, friends, biography, education, work history etc.), posts (wall posts, including images, comments, locations, etc.), news feeds (data from other users, including images, comments, locations etc.), check-ins (places, comments, applications etc.), events (venues, times, locations, descriptions, participants, etc.), groups, and status updates (times, messages, comments, etc.).

Researchers can check a wide range of data types in our system, and we intend to provide access to the majority of the data, but the researcher can only choose the data type needed for his/her specific research project.

Data analysis
When the data is collected, Digital Footprints facilitates different kinds of searching and sorting options. To support the opportunity for researchers to study both individual users and user patterns across users in the panel, the system builds on two sorting mechanisms (panel sorting and user sorting).

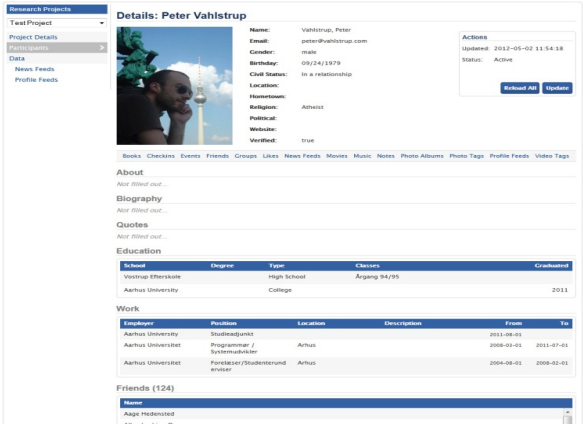

Illustration 1. Stem data for a single participant in a research project.

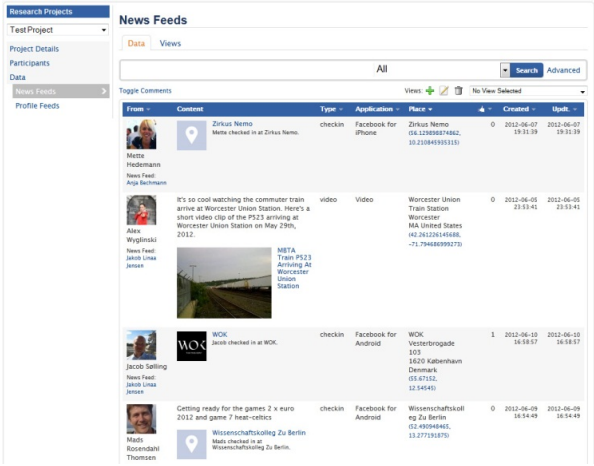

Illustration 2. News feeds across participants in a particular research project.

For example, researchers can either see wall posts for all users participating in a research project, and search the wall posts for keywords, or see the wall posts of one specific participant. It is possible to make ranged queries by defining start and end dates. The current version also supports descending and ascending sorting of categories such as "likes", "location", "application", "birthday", and "civil status", along with option of creating different "views". The 'views' function allows the researcher to select specific data units and transfer them to different sub-projects.

Findings and discussion

The design of the multiuser data retrieval app Digital Footprints and the user feedback from researchers internationally do show optimization of the data access in terms of creating searchable and sortable data, accessing and retrieving historical Facebook data and personal data views such as the newsfeed, being more unobtrusive than observation studies, and the app do not require programming skills. However, the multiuser app design and the use of open APIs in qualitative studies of social media interaction and communication in general raise new methodological challenges. In this section we bring forward three of the main issues.

Interface differences

In order to secure the reliability and transparency in qualitative studies the data received through the app need to match the data view of the research participants [19]. Neuhaus & Webmoor [22] argue that the public API of Twitter only gives access to a sample size of the actual data streams. Our design shows that this seems not the case with Facebook. However, data collection is heavily dependent on Facebook servers. For instance, using Digital Footprints in the first qualitative user study we had an empty time frame in the profile feed data. This was most likely due to a server outage at Facebook, and we had to restore the data by re-retrieving it from Facebook at a later date.

Through Digital Footprints, our research project was very dependent on the Facebook categories. For example, the research project was designed to investigate how many and what kinds of apps the participants had permitted to draw data, but the category did not exist in the form of data that developers could draw from Facebook, through its open API. Also, we wanted to determine how many posts were updated from a third party application such as Twitter, Spotify, Instagram, or YouTube. The open API supports this functionality, but when analyzing the data, no entries from Spotify and Pinterest, for example, appeared, even though we could see from the screen dumps that participants had connected to these apps. Facebook's open API is not well documented, but we suspect that the missing entries are due to third party apps using the actions API. To get hold of these entries we need to request permissions to access each individual namespace for the apps used. To some research projects the request is difficult because they have to know which apps the user has installed in order to automate the information retrieval. For other projects focusing on a particular app (e.g. Spotify) this is trivial as one can assume that all users have the app and therefore retrieval can easily be automated.

Finally, we wanted to register the news feed flow to which participants were exposed. However, the data showed that the open API transferred all entries to which the participants could potentially be exposed. Surprisingly, Facebook does not register whether or not

a user has been exposed to an entry. This means that as methodological tool, Digital Footprints cannot simulate the actual flow on the user's screen, because we do not know whether the user has actually been exposed to all updates. For example, this would not be the case if s/he logs in to Facebook very seldom, or has many friends/potential updates. In these cases, Digital Footprints retrieves more updates than the user has been exposed to. On the positive side, the open API allowed the research project to use exact time stamps on each entry, the exact coordinates for location-tagged data, and we were able to contextualize each data unit, because each unit has a unique identification code.

Legal and ethical issues

When securing the informed consent of participants, we draw attention to the fact that researchers are bound by privacy regulations to anonymize data collected from users [9,16,24], but we also make sure to state to the end-user that several cases have shown that it is possible to de-anonymize data and text-strings [16,24]. Even though we try to follow the basic privacy by design principle of transparency [14] we are caught by the privacy as well as transparency paradox [23]. Users (including our research participants) neglect to read privacy policies and if transparency becomes too complex, as is the case with a large number of permissions on Facebook, users do not read through the permissions. How do we secure that user are informed then? We encourage researchers to follow the three-step user consent process thoroughly.

Even though qualitative researchers always have worked with personal and potential sensitive data many researchers was concerned about the fact that the researcher can assess content posted by participants' friends exposed through the participant. One of our test researchers phrased it this way: "*you were researching one person, and suddenly a lot of other people are popping up*". This is not a legal matter as Facebook owns the right to data and data-sharing with third party companies, but it certainly raises ethical concerns. Is it okay to study participants' friends as we qualitative researchers would have done during observation studies if they came into a physical setting or should we anonymize friends in the app so that researchers do not see their names but only the content posted? If we choose the latter, how do we anonymize photos uploaded by friends or picturing friends? These considerations must be held against studies on social media and privacy arguing that social media users want to share their data and know the premises for data sharing with third parties [e.g. 2]

Holding workshops in different countries, made us realize that there are very different attitudes towards and approaches to applying for permission to carry out research projects that access personal, potentially sensitive information. Even in Scandinavia, the authorities/Data Protection Agencies have very different procedures, and demand different things of the researchers even though we operate under EU law. Even though United States has a very liberal approach to privacy the IRBs have strict rules for qualitative research, following yet other procedures. Designing an international multiuser app makes this difficult. How should we approach this ethically as well as legally? And does this issue points to a more substantial and general question that needs to be addressed on a higher level?

Conditions of use
Another principal matter is Facebook's conditions of use. When users created a Facebook account, they agreed that Facebook owns all their data and traffic patterns. Therefore, we and other researchers using the app need to follow Facebook's developer guidelines, when drawing user data from Facebook and the app depends on the willingness of Facebook to keep (nearly all) data public available.

During the spring 2012 we contacted Facebook to let them know about the app that we were designing and to initiate a more principal discussion on using their open API for research purposes. The aim of the open API is to enhance content and user engagement in the Facebook data sphere. How is a data retrieval app social and engaging? We argue that participants engage in the research project and by participating in the research project and sharing their data they increase their user experience of Facebook in a more informed way. The idea is to develop the app so that participants can see their own data, and thereby see what they sent to (all) third party apps on Facebook, when they more or less consciously click "allow".

Conclusion & future direction

Through the design of Digital Footprints, an app for Facebook data retrieval, we have shown how qualitative user studies can benefit from the use of the open API. Based on existing methods, the system design, and the user iterations, Digital Footprints shows a great deal of promise for improving the efficiency of data collection, and the validity of data because it is possible to collect all data (content units) from the participants. Similarly, in this article we have shown that the software has great potential for lowering researcher workload, and maintaining focus on the research topic at hand, when analyzing social media data.

At the same time the paper stresses that working with the API can be methodologically problematic at times, especially with concerns to interface differences. Furthermore, the paper shows how cultural differences play a major role when addressing necessary legal and ethical issues. This has made us realize that we need to facilitate the opportunity for researchers to write about their experiences with country-specific approaches to research permissions.

It is important to not be over-enthusiastic about the opportunity to collect data from open APIs, and the ease with which this may be done. Researchers need a clear purpose when doing this, because Digital Footprints may deliver data, answering the question, "What?", but it does not often answer the question, "Why?", in qualitative research. We need to consider whether it is possible to facilitate closer contact with the participants in a research project during the data retrieval period, to simultaneously address the "why?" questions. This may be done either through the Facebook inbox, app or texting, perhaps connecting a notification to researchers when adding certain types of data.

ACKNOWLEDGMENTS

We thank Aarhus University (DAC) and Digital Humanities Lab (DK) for funding the system design and for support on the legal matters of our app. Furthermore we thank Facebook for engaging in a dialogue on the terms of use interpreted in this specific research context.

REFERENCES

[1] Baym, N.K. *Tune in, Log on. Soaps, Fandom, and Online Community*. Thousand Oaks, CA: Sage Publications, (2000).

[2] Boyd, D. & Marwick, A. Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies, *Decade in Internet Time Symposium*, OII (2011).

[3] Brandt, J., Weiss, N. and Klemmer, S.R. txt 4 l8r: lowering the burden for diary studies under mobile conditions. In *CHI EA 2007*, ACM, (2007).

[4] Bruns, A. How Long Is a Tweet? Mapping Dynamic Conversation Networks on Twitter Using Gawk and Gephi. *Information, Communication & Society*, Nov. 17 (2011).

[5] Cantanese, S.A., Pasquale..D.M., Ferrara, E., Fiumara, G. & Provetti, A. Crawling Facebook for Social Network Analysis Purposes, *WIMS2011*, ACM, (2011).

[6] Carter, S. and Mankoff, J. When Participants do the Capturing: The Role of Media in Diary Studies, *Proceedings of CHI2005*, (2005), pp. 899-908.

[7] Consolvo, S. and Walker, M. Using the Experience Sampling Method to Evaluate Ubicomp Applications. *IEEE Pervasive Computing* 2, 2 (2003), p. 24-31.

[8] Courtois, C., Mechant, P. & Marez, L. Teenage uploaders on YouTube: Networked public expectancies, online feedback preference, and received on-platform feedback, Cyberpsychology, Behavior, and Social Networking, vol. 14(5), (2011),

[9] *Directive 95/46/EC*, Official Journal L 281, 23/11/1995, p. 31 – 50.

[10] *Directive 2002/58/EC* (on privacy and electronic communications), L201/37.

[11] Gjoka, M., Sirivianos, M., Markopoulou, A. and Yang X. Poking Facebook: Characterization of OSN applications, *Proceedings of the first workshop on online social networks*, ACM, (2008), p. 31-36.

[12] Hammersley, M. and Atkinson, P. *Ethnography: Principles in Practice*, Routledge, (1995).

[13] Hine, C. *Virtual Ethnography*, Sage, (2000).

[14] Kelley, P.G., Cesca, L., Bresee, J. and Cranor, L.F. Intentional Design of Privacy Policies, *CHI 2011 Workshop on Networked Privacy*, ACM, (2002).

[15] Kendall, L. *Hanging Out in the Virtual Pub: Masculinities and Relationships Online*, University of California Press, (2002).

[16] Krishnamurthy, B. and Wills, C.E. On the Leakage of Personally Identifiable Information Via Online Social Networks, *WOSN2009*, (2009), pp. 7-12.

[17] Korn, M. From Hybrid Spaces to Experiencing Augmented Places, presented at the conference *Mobile Communication: Mobile Internet, Locative Media, Mobility and Place*, March 29-30, (2012).

[18] Kramer, A. The spread of emotion via Facebook, *CHI2012*, ACM Press, (2012), pp. 767-770.

[19] Kvale, S. & Brinkmann, S. *Interviews: Learning* the craft of qualitative research interviewing, Sage, (2008).

[20] Magnani, M, Montesi, D. and Rossi, L. Conversation Retrieval from MicroBlogging Sites, *Information Retrieval Journal*, Volume 15, Issue 3-4, (2012), pp 354-372.

[21] Markham, A.N. *Life Online. Researching real experience in virtual space*, AltaMira Press, (1998).

[22] Neuhaus, F. & Webmoor, T. Agile ethics for massified research and visualization. *Information, Communication & Society* 15(1), (2012), pp. 43-65.

[23] Nissenbaum, H. A Contextual Approach to Privacy Online, *Daedalus*, 140 (4), (2011), pp. 32-48.

[24] Ohm, P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Review, Vol. 57, (2010), p. 1701.