Data Collection for Internet Researchers under Data Protection Regimes

Compiled by Seamus Allen
April 2023

IIEA
Sharing Ideas
Shaping Policy

MARQUETTE UNIVERSITY
Center for Data, Ethics, and Society

Together with
Meta

A conference hosted by Meta Ireland in partnership with the Center for Data, Ethics, and Society, Marquette University and the Institute of International and European Affairs (IIEA), Tuesday, 1 November 2022, Serpentine Avenue, Ballsbridge, Dublin.

Session 1: A Panel on the Future of Data Sharing

This panel was composed of:

- **Dr. Rebekah Tromble**, Associate Professor in the School of Media and Public Affairs and Director of the Institute for Data, Democracy, and Politics at George Washington University;
- **Dr. Richard Rogers**, Professor of New Media and Digital Culture at the University of Amsterdam;
- **Bennett Hillenbrand**, Privacy and Data Policy Manager; Research Data Sharing and Academic Transparency at Meta;

and was moderated by:

Dr. Michael Zimmer, Associate Professor of Computer Science and Director of the Center for Data, Ethics, and Society at Marquette University

In this panel discussion, a number of key themes relevant to data access for researchers were explored, including:

i) the General Data Protection Regulation (GDPR);
ii) new initiatives and policies;
iii) technical and practical challenges confronting researchers, and;
iv) the relationship between researchers, platforms, and other stakeholders.

At the start of the panel discussion there was clarification given regarding the status of access to 'CrowdTangle' for researchers and about future communications regarding the tool.

A major part of the discussion centred on the GDPR and the actual legal implications of this regulation versus the regulation's perceived implications. It was generally agreed that the GDPR does indeed impose some major restrictions on data access for researchers. However, the point was also made that in some cases the GDPR was improperly cited as a reason for restricting researchers' access to data, even when such restrictions were not required by the regulation. A view was also expressed that sometimes researchers themselves did not have a fully clear understanding of their need to accept responsibilities and liabilities under the GDPR.

The discussion also examined some recent initiatives and developments, such as the Digital Services Act (DSA), which will impose obligations on online platforms relating to data access for researchers. It was suggested that the European Digital Media Observatory (EDMO) Code of Practice on access to data for researchers could be used as a possible blueprint for a future delegated act under the DSA relating to data access for researchers. There was also discussion about the role of Digital Service coordinators created under the auspices of the DSA. It was argued that a lot of political work might be required to convince Digital Service coordinators to enforce data access for researchers. These coordinators may be able to improve data access for researchers in a variety of cases, but it was also suggested that such improvements would only occur in some situations and "bottlenecks" were expected to occur. It was therefore argued that the DSA would not open the "floodgates" for data access. There was also discussion regarding the establishment of an EDMO working group 2.0 and a new intermediary body that would be created which would be independent of both governments and platforms. This intermediary body would monitor the data access that platforms give to researchers and the compliance of researchers with ethical requirements. It was stated than an intermediary body would be established by EDMO in the next 12 months and that there may be pilot projects during this period. This intermediary body will aim to ensure that both platforms and researchers are doing the right thing.

The discussion also explored some of the practical and technical challenges faced by researchers when attempting to access relevant data. One example was the difference between "developer mode" and "user mode" access. Researchers often want to see specifically what users see, but their mode of access may render the platform in a different way. Another challenge discussed was that of "missing data" – data that had been either removed or deleted. For example, this could occur when networks of disinformation were removed and researchers might thus lose access to valuable data. A further challenge concerns non-sensitive data relating to sensitive topics or that is related in some way to other more sensitive data. It was pointed out that particular queries could pose major problems when it comes to accessing content, even if such queries were 'adjacent' to other topics that did not pose such problems. Others argued that there are different categories of personal data and risk, and that even research into non-sensitive data could be high risk if it could be used to infer other types of information, e.g. inferring political beliefs from non-political data. Thus, it was argued that non-sensitive data could be used to produce sensitive outputs. However, the view was also expressed that even research into sensitive topics should not necessarily be deemed to be high-risk, as it would not necessarily use sensitive data (for example if the data used was entirely aggregated and anonymised). It was argued that in some cases access to data for sensitive topics was restricted by platforms even when the data itself was not sensitive or personal in nature. A member of the audience noted that people belonging to certain ethnic communities can face greater difficulty in researching their own ethnic groups due to this being regarded as sensitive data. This increases bureaucracy and barriers to undertaking research about issues confronting these communities. There were different views expressed with regards to whether 'differential privacy' was desirable. It was noted that the EDMO working group did not recommend differential privacy. It was also suggested in the discussion that a multiplicity of models regarding data access for researchers was required and that no single approach would solve everything.

The relationships between researchers and platforms, and between researchers and other stakeholders also comprised part of the discussion. Here participants emphasised the importance of shared accountability between platforms and researchers which should include safeguards and oversight for both researchers and platforms. Participants also discussed the need for researchers to be accountable to the people that they study. A viewpoint was expressed that if accountability or liability for the work of researchers was held by a third party, the independence of researchers may then be called into question. It was also argued that some parts of the research community can sometimes do itself a disservice when it comes to topics relating to accountability and it was suggested that not all researchers do things for the right reasons. The view was also expressed that workforces in some platform companies were disproportionately from certain types of demographic groups, potentially leading to policies that are less advantageous for other demographic groups. It was separately noted that the GDPR gave no opt-out to journalists unlike the 'carve out' it gave to researchers.

Overall, the conversation in this panel focused on the ethics, regulations and policies relating to data access for researchers. There was particular interest shown by participants in how best practice for data access can be established and how unnecessary obstacles to data access might be overcome.

Session 2: A Panel on Balancing Community Value with Individual Choice

This panel included:

- **Dr Nicholas Proferes**, Assistant Professor in Arizona State University School of Social and Behavioural Sciences;
- **Dr Sophie Bishop**, Lecturer in Sheffield University Management School;
- **Dr Jennifer Stromer-Galley**, Professor in Syracuse University School of Information Studies;

and was moderated by:

Professor Joyce O' Connor, Chair of the IIEA Digital Group

The theme of ethical obligations for researchers, influencers, and platforms was discussed in this conversation. The discussion centred particularly on challenges relating to ethical obligations for researchers. As an example, there was a discussion about how it could be difficult for researchers to determine if the communities that they researched counted as private communities or as public communities. There was also discussion about the rights for "malfeasant actors" that researchers may wish to study, and whether such actors had rights to not be studied. The distinction between legal and ethical obligations was particularly discussed. It was noted that social media users often have inaccurate mental models or "folk models" of how online platforms actually work and this may affect the significance of user consent. In the discussion it was suggested that in cases where informed consent was not possible other ethical options should be considered. Such options might include opt-outs, debriefing, and sharing results. There was also some debate about where and when consent is required. The view was expressed that wherever there is experimentation on users of a platform in order to obtain knowledge, consent should be necessary. The way that online databases can be misused by researchers was noted, with an example provided being that of classification schemes which organise and label pictures without the knowledge of users. It was noted that there is no ‘one size fits’ all policy that can fix or solve everything in all cases, but that an ethic of care should underpin the approach taken by researchers. This ethic of care must also apply to “hostile” communities.

There was also discussion about the ethical responsibilities of influencers, and the ethical responsibilities of platforms towards influencers. It was emphasised that influencers have ethical obligations to be transparent and honest with their followers and communities. The discussion also examined how social influencers or businesses who use platforms can be economically jeopardised by the decisions made by platforms – including the banning or restricting of content. This implies that there are ethical responsibilities that platforms should take into account when setting policies that affect influencers.

Overall, the conversation in this panel focused on the ethical obligations of researchers, influencers, and platforms. There was particular interest expressed by participants in topics such as whether consent is always necessary or possible to obtain and regarding the ethical alternatives to consent. There was also particular interest shown in whether ethical obligations of researchers might differ between different types of communities that they may study.

Session 3: Break-out Discussion Groups

To undertake a ‘deeper dive’ into the question of how the correct balance could be struck between the “Community Value” created by data access for academic research and the choice and control individuals have over their own data, the room was broken into discussion groups, with 5-8 participants in each group. Two hypothetical scenarios were presented to facilitate this exploration. In one scenario, researchers were asked to reflect on the approach that might be taken to researching a large AI system, like a Newsfeed or Recommender system. In a second scenario, participants reflected on tools to provide transparency into an advertising system.

The groups were prompted to imagine what an ideal set of processes and tools might look like to provide transparency into each system. They were then asked to consider the user experience – what level of transparency, consent, or choice should a person have over their data being used by academics for this research?
A number of subjects were discussed within the groups relating to themes such as consent and alternatives, the differentiation of users and data, and features of platforms that pose challenges for researchers. One important theme related to consent, notice, and opt-outs for data subjects. One group stated that to pursue the user journey researchers needed to be able to account for personalisation effects. For example, the importance of assessing what factors influence algorithms was emphasised as it might be necessary to know how the user experience would change if any one variable were to be changed. It was argued that there must be a way for a user to opt-out from, and to be notified about, such research. It was suggested that it could be useful to explain the value proposition to users regarding the benefits of the research being undertaken. Another group was of the view that consent would not be required if only aggregated data is used. It was also argued that giving notice to users and providing opt-outs can in fact skew

responses and thus affect the quality of research findings. A separate question was raised in relation to the ethical implications of situations in which there is a difference in consent between a parent and a minor, and how such a situation should be addressed.

Another theme of the discussion focused on the ability to differentiate between types of users and types of data based on sensitivity and risk. The view was expressed that there should be different levels of sensitivity when handling different types of data. The importance of a risk-based approach was noted, with the implementation of transparency and consent systems based upon the level of risk. The importance of determining how the platforms categorise and differentiate their users for the purposes of targeting was also discussed.

Challenges relating to policies or features of online platforms themselves were also discussed. There was some discussion about how it would be valuable for archives to be maintained for future scrutiny, as the value of certain types of research may only be appreciated in the future. For example, the harmfulness of certain advertising may only become apparent years later. One group argued that data explicitly provided by platforms might not be the same thing as the data that is actually used by other platform stakeholders that a researcher is investigating within the same research context. For example, advertisers may use data provided by Meta to try and 'microtarget' users, even if Meta itself was not seeking to explicitly microtarget its own users. Therefore, if researchers simply use the data that is directly provided by the platform, they may not be studying how this data may actually be ultimately used by advertisers. Another view expressed was that lawyers representing platforms should not be the ones deciding what data researchers can access. It was argued that an intermediary body that undertakes ethical reviews must be involved with the more difficult calls.

The discussion in the breakout session ultimately focused on the need to balance the possible benefits of using personal data for research with the need to empower individuals with choice and control over how their data might be used. There was particular interest shown in the importance, the potential disadvantages, and alternatives to obtaining consent, the differentiation between different types of users and data, and the features or policies of platforms that pose challenges to researchers.

Concluding Thoughts

Throughout the day, discussions and contributions ranged across matters concerning the ethics, regulation, and policies relating to data access for researchers, among much else. There was particular interest shown by participants in how best practices for data access can be established and how unnecessary obstacles to data access can be resolved. There was considerable interest expressed in how to determine the scope of when consent was necessary. The importance of drawbacks of seeking consent was also discussed as well as the existence of ethical alternatives to seeking user consent. There was also considerable interest expressed in how researchers should differentiate between the different types of users, communities, and personal data that might relate to a study, depending on factors such as risk, sensitivity, and privacy.

The Institute of International and European Affairs (IIEA) is Ireland's leading international affairs think tank. Founded in 1991, its mission is to foster and shape political, policy and public discourse in order to broaden awareness of international and European issues in Ireland and contribute to more informed strategic decisions by political, business and civil society leaders.

The IIEA is independent of government and all political parties and is a not-for profit organisation with charitable status. In January 2021, the Global Go To Think Tank Index ranked the IIEA as Ireland's top think tank.

© Institute of International and European Affairs, April 2023
Creative Commons License
This is a human-readable summary of (and not a substitute for) the license.
https://creativecommons.org/licenses/Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
You are free to:
• Share - copy and redistribute the material in any medium or format
• Adapt - remix, transform, and build upon the material
• The licensor cannot revoke these freedoms as long as you follow the license terms.
Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
NonCommercial — You may not use the material for commercial purposes.
ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.
No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

The IIEA acknowledges the support of the Citizens, Equality, Rights and Values (CERV) Programme of the European Union

IIEA
Sharing Ideas
Shaping Policy