«Comunicazioni sociali», 2023, n. 2, 139-151
Published online October 2023

ELISABETTA RISI - GUIDO DI FRAIA*

COLLECT AND HANDLE PERSONAL SOCIAL MEDIA DATA
Ethical Issues of an Empirical Internet Research

Abstract
Online users' digital traces provide valuable information and empirical evidence, but Internet research requires scientific rigor in accessing and managing User Generated Contents (UGCs). The article challenges these practices and advocates for a reflexive approach to social media research ethics. Although platforms offer viable access, utilizing such data can intrude on subjects' private lives. Defining responsibilities toward data and subjects is crucial when studying online contents, such as Instagram stories and Facebook posts. The subject's centrality and ethical implications becomes particularly significant in social inquiry, where the object is closely tied to actively signifying subjects and social relations mediated by institutions or technologies. The paper explores ethical issues in a concrete research project, "7 friends for 7 days", and presents alternative research practices for observing and analyzing online content within the post-API research context. It discusses ethical challenges in Internet research, focusing on social media data, and examines a study that analyzed user-generated content through human-type coding. The paper reflects on the ethical considerations in fabricating research evidence, particularly regarding UGC published on personal social media platforms and the critical awareness of those involved in observing and disseminating such data.

Keywords
Social media research ethics; research practices; User Generated Content; ethical implications.
ISSN: 03928667 (print) 18277969 (digital)
DOI: 10.26350/001200_000181

1. INTRODUCTION. THE ESSENTIAL PRESENCE OF ETHICS IN RESEARCH

While the ethical question has consistently played a significant role in social research[1], treating the Internet realm as a tool, field and venue for investigation requires an *ad-hoc* and far-reaching ethical discussion[2]. Specifically, the current ubiquity and pervasiveness of digital technologies demands that the standards for the protection of users' data should be raised. Social media have accelerated the process of "datafication" of soci-

* IULM University – elisabetta.risi@iulm.it; IULM University – guido.difraia@iulm.it

[1] G. Di Fraia, "L'etica nelle ricerche di mercato", in *Studi di Sociologia*, 48, 3-4 (2010): 309-321.

[2] A growing number of scholars have explored these questions and scholarly associations have stated some fundamental ethical guidelines for Internet research (see, among others, C.M. Ess, the AoIR Ethics Working Committee, *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee*, 2002); a.s. franzke, A. Bechmann, M. Zimmer, C.M. Ess, the Association of Internet Researchers, *Internet Research: Ethical Guidelines 3.0*, 2020, from https://aoir.org/reports/ethics3.pdf.

ety[3], as well as being suitable for studying social phenomena. As a result, the web has become the repository of our *onlife* lives[4], and the contents daily published by people are being monitored by at least three agencies: first, inter-connected publics[5] reciprocally checking their own profiles; second, researchers who analyze and sociologically interpret these digital traces and footprints[6]; third, individuals are surveilled by digital platforms that continuously feed on these data, thus contributing to the livelihood of the platform capitalist system[7].

Even though online users' digital traces constitute a rich mine of information and empirical evidence that is sometimes inaccessible otherwise, their usage should require a case-by-case approach, taking into account the implication for ethics[8]. While for instance social research based on user generated contents (here and after, UGCs) could also be implemented through experimental and creative attempts[9], Internet research requires scientific rigor, in both access and management of contents of/on subjects.

The Internet is a tempting place where one can gather information, which is populated by many "sirens"[10], who swim in the great basin of contents published and shared online. This scenario has posed both opportunities and challenges to social researchers. Like sirens, these allegedly free, and available sets of information could make scholars lose their ethical bearings along the way.

Thus, assuming that social media users have knowingly posted "public domain" content, this "opt-out" approach appeared to have paved the way for the wide-ranging collection of UGCs used as empirical data in social research. Accordingly, in this article, we will problematize such practices, using a more reflexive approach to social media research ethics. Although platforms affordances-wise[11] viable, the access to these data, has made available information whose fruition would constitute, by its nature, an intrusion into the private life of the subjects. Therefore, in developing methods for studying the types of UGCs (such as Instagram storied and Facebook posts) most published and shared, we point to the importance of defining responsibilities[12] with respect to both data and subjects involved in research. When it comes to online content, there is therefore an intertwining link between subjects (users) and data (contents).

The digital footprints left by online users represent a valuable source of data that may be otherwise unavailable. However, their utilization should be approached on a case-by-case basis, considering on the one hand the centrality of the subject and on the

[3] E. Risi, *Vite datificate: modelli di ricerca nella società delle piattaforme*, Milano: FrancoAngeli, 2021.

[4] L. Floridi, *La quarta rivoluzione: Come l'infosfera sta trasformando il mondo*, Milano: Raffaello Cortina Editore, 2017.

[5] Z. Papacharissi, "On Networked Publics and Private Spheres in Social Media", in *The Social Media Handbook*, edited by J. Hunsinger and T. Senft, Andover: Routledge, 2013, 144-158.

[6] F. Comunello, F. Martire, L. Sabetta, eds., *What People Leave Behind. Marks, Traces, Footprints and their Relevance to Knowledge Society*, Cham Switzerland: Springer, 2022.

[7] N. Srnicek, *Platform Capitalism*, Cambridge: Polity Press, 2017.

[8] M. Ananny, "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness", *Science, Technology, & Human Values*, 41, 1 (2016): 93-117.

[9] L. Bainotti, A. Caliandro, A. Gandini, "From Archive Cultures to Ephemeral Content, and Back: Studying Instagram Stories with Digital Methods", *New Media & Society*, 23, 12 (2021): 3656-3676. DOI: 1461444820960071.

[10] C. Cipolla, A. De Lillo, E. Ruspini, *Il sociologo, le sirene e gli avatar*, Milano: Franco Angeli, 2012.

[11] T. Bucher, A. Helmond, "The Affordances of Social Media Platforms", in *The Sage Handbook of Social Media*, edited by J. Burgess, A. Marwick, T. Poell, London: Sage, 2017, 254-278.

[12] C. Sandvig *et al*., "When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software", *International Journal of Communication*, 10 (2016): 4972-4990

other the more general ethical implications involved. To avoid the use of ad hoc rules, this approach must in fact be balanced with dialogue with the ethical guidelines.

In our view, the ethical question is fundamental not simply to assess the moral compass of research conduct, but to be self-reflexive of the ontological and epistemological assumptions that guide such a conduct. In this case, we push against the cartesian/post-positivist assumption that distinguishes "observing subject" and "observed object", assume instead their dialectical and reciprocal constitution. This subject-object dialectics becomes even more significant when we consider that in social inquiry, the so-called "object", as a social phenomenon, is actually mostly tied to actively signifying subjects and/to social relations among subjects mediated by institutions or technologies, which means that, while maybe not direct, the interaction when we deal with "observing subject"-to- "observed subject" relations is possibly even superior.

This sort of anti-cartesianism builds on Markham's idea[13] that both methods and ethics are strengthened conceptually and practically when researchers impose the characteristics and functions of each concept onto the other. This is not an easy task because the locus of responsibility and accountability for ethical design, behavior, and outcomes is difficult to ascertain. To this purpose, we will explore those ethical issues associated to a concrete research project, namely "7 friends for 7 days"[14], in order to illustrate how we have developed a set of practices to ensure that this research was ethically responsible.

This article focuses on Instagram and Facebook, which are platforms with a specific and limited content/data access premises. We will therefore suggest an alternative form of research practices for the observation and analysis of online content in the context of the so-called post-API research[15].

Thus, this article discusses some ethical challenges of Internet research with a specific focus on social media data and, after an overview of the debate around these issues, it presents the case of a recent research project based on a collection of users generated content that was treated as relatively contained kind of data, and therefore can be analyzed through human-type coding[16].

Drawing on a critical examination of methodological choices made in our empirical study, this paper reflects on the ethical fabrication[17] of the research evidence. These reflections concern, on the one hand, the fact that the digital data analyzed are UGCs published on personal social media and therefore constitute the daily storytelling of the observed subjects[18]; on the other hand, it is examined the critical awareness of those who have observed and coded such data with respect to the ethical issues of this research practice.

[13] A. Markham, "Ethic as Method, Method as Ethic: A case for reflexivity", *Journal of Information Ethics*, 15, 2 (2006): 37-54. DOI: 10.3172/JIE.15.2.37

[14] R. Pronzato, E. Risi, "7 Amici per 7 Giorni. Uno studio sulle tracce digitali degli utenti su Instagram e Facebook", in *Tracce digitali e ricerca sociologica*, a cura di E. Risi e A. Gandini, Milano: Franco Angeli, 2023, 85-104.

[15] D. Freelon, "Computational Research in the Post-API Age", *Political Communication*, 35, 4 (2018): 665-668. DOI: 10.1080/10584609.2018.1477506; A. Bruns, "After the "APIcalypse": Social media platforms and their fight against critical scholarly research", *Information, Communication & Society*, 22 (2019): 1544-1566. DOI 10.1080/1369118X.2019.1637447.

[16] M. Luka, M. Millette, "(Re) Framing Big Data: Activation Situated Knowledge and a Feminist Ethics of Care in Social Media Research", *Social Media & Society*, 4, 2 (2018): 1-18. DOI: https://doi.org/10.1177/2056305118768297

[17] A. Markham, "Fabrication as Ethical Practice: Qualitative Inquiry in Ambiguous Internet Contexts", *Information, Communication & Society*, 15, 3 (2012): 334-353.

[18] These aspects are part of the broad theme (which we will not go into further) of the ways that algorithms co-construct narrative identity and relational meaning in contemporary use of social media, especially for young people.

2. NO DATA ARE CREATED EQUAL: MAIN ETHICAL NODES OF SOCIAL MEDIA DATA

The buzzword *Big Data* indicates the large amount of data produced, stored, and analyzed by digital platforms. In addition to their volume, big data are also characterized by other features[19], which make them capable of offering detailed and large-scale information on the relationships and to the individuals who constitute them. However, in the process of interfacing with these sources, critical issues may arise, and we argue that the academic world cannot underestimate them[20]. Firstly, those data rest on digital platforms and apps managed by a group of technology giants or tycoons (e.g. Meta, Amazon etc.), that follow first and foremost an economic logic in handling data access necessary to maintain an eco-system of apps and services that benefit the platform itself [21].Second, just because those data sets are big doesn't mean they are complete, representative, neutral or accurate. While digital platforms like social media allow the retrieval of remarkable amount of data, each dataset has limitations and biases[22].

Another issue is tied to the automated systems used to collect data. Possibly, the most widely used method is the data downloading technique through query tools provided by the so-called APIs[23].

In social research, scholars mainly deal with *unstructured data*, i.e. information of a textual, iconic or video nature, which would require various preliminary processing such as data mining (i.e. systematization, labeling and coding treatments), necessary to transform this raw information into usable data for statistical analysis. Dealing with this un-structuredness, the social researcher can also pursue a *small data* approach, mainly qualitative and ethnographic logics[24] to collect detailed information on users lives. Small data are composed of a relatively small collection of datapoints or cases, so that their analysis can be performed single-handedly via human coding.

Both small and big data must be guided by ethical principles. Fundamental principles of conducting ethical social research remain the same[25], but in the face of the technological mediation of the platforms and the 'quantity' of data available to the researcher, Internet research implies some specific ethical dilemmas[26]. The fact that "just because it's public it doesn't mean it can be freely used for research" epitomizes how the literature dealing with social media ethical issues has exploded during the last 15 years.

[19] R. Kitchin, G. McArdle, "What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets", *Big Data & Society*, 3, 1 (2016). https://doi.org/10.1177/2053951.

[20] N. Marres, *Digital Sociology: The Reinvention of Social Research*, Cambridge: Polity Press, 2017.

[21] D. Trezza, "To Scrape or Not to Scrape, This Is Dilemma. The Post-API Scenario and Implications on Digital Research", *Frontiers in Sociology*, 8: 1-8 (2023). DOI: https://doi.org/10.3389/fsoc.2023.1145038.

[22] d. Boyd, K. Crawford, "Critical Questions for Big Data. Provocations for a Cultural, Technological, and Scholarly Phenomenon", *Information, Communication & Society*, 15, 5 (2012): 662-679.

[23] Standing for Application Programming Interface (API), this acronym indicates the set of procedures and programs that allow the researcher to make a request for data acquisition to a digital platform (precisely through API), extracting the data directly from the server. However, this path faces limitations, because each privately-owned platform establishes very specific constraints on the type and quantity of data that can be obtained.

[24] A. Caliandro, A. Gandini, *Qualitative Research in Digital Environments: A Research Toolkit*, London: Routledge, 2016.

[25] Ess and the AoIR Ethics Working Committee, *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee*.

[26] S. Livingstone, E. Locatelli, "Ethical Dilemmas in Qualitative Research with Youth On/Offline", *International Journal of Learning and Media*", 4, 2 (2014): 67-75. DOI: 10.1162/IJLM_a_00096.

Danah boyd, pioneered ethically-aware research practices, mainly by questioning how privacy was linked to the awareness that users had on which information was shared with whom, and in what context[27].

The possibility of using publicly accessible content/data is therefore limited in compliance with fundamental regulatory concepts such as that of privacy, and the boundaries between public and private. However, far beyond the legal provisions on personal data, the acquisition of data through digital platforms requires the researcher to exercise caution in managing the digital 'traces' of the subjects/users.

Social science research on Facebook and Instagram has identified several ethical challenges related to the collection, analysis and reporting of data[28]. Even before the Facebook–Cambridge Analytica scandal of 2018, these platforms had in fact introduced important settings that allowed users to select the audience whose the content they publish would be visible to, and therefore actively decide what to make available to the "public", rather than sharing with only a small group of "friends", or only "privately" with a few selected users. For example, the direct message function (2013) allows users to share contents only with specific contacts[29], while the "Stories" function (2016) was added to allow users to post content visible only for 24 hours (unless saved to *Stories IG Highlights* or republished as a post).

Firstly, researchers could in principle rely on these settings to identify what is "publicly available content". Furthermore, these same accounts can be hooked up via APIs, which allowed big data downloading operations. Personal profiles, typically considered to be information-rich for digital sociological research, are excluded. So, as we will see in the next paragraph, the study of the contents published on these accounts can only take place with post-API research approaches[30].

While informed consent provided by users would appear to be unnecessary for the use of content posted on public accounts, public availability is not sufficient for ethically sound academic research[31]. That is because sometimes terms and conditions signed upon registration are not clear-enough to define whether a content can be used for research purposes or not.

Moreover, users have different awareness of how their data are viewed and/or consulted by so-called third parties, including scholars[32]. Thus, especially considering the data-subject interweaving perspective advanced here, scholars should act through informed consent which does not only aim at defending privacy but also at stimulating critical awareness of the use of digital platforms and implementing effective accountability mechanisms[33].

There are today guidelines on many of these aspects mentioned above, including

[27] d. boyd, "Facebook's Privacy Trainwreck: Exposure, Invasion and Social Convergence", *Convergence*, 14 (2008): 13-20.

[28] E. Locatelli, "Images of Breastfeeding on Instagram: Self-Representation, Publicness, and Pivacy Management", *Social Media+ Society*, 3, 2 (2017): 1-14. DOI:10.1177/2056305117707190.

[29] T. Highfield, T. Leaver, "A Methodology for Mapping Instagram Hashtags", *First Monday*, 20, 1 (2015). https://doi.org/10.5210/fm.v20i1.5563.

[30] Bruns, *After the "APIcalypse"*.

[31] K. Tiidenberg, "Ethics in Digital Research", in *The Sage Handbook of Qualitative Data Collection*, edited by U. Flick, London: Sage, 2018, 444.

[32] S. Ravn, A. Barnwell, N.B. Barbosa, "What Is "Publicly Available Data"? Exploring Blurred Public-Private Boundaries and Ethical Practices through a Case Study on Instagram", *Journal of Empirical Research on Human Research Ethics*, 15, 1-2 (2020): 40-45. DOI:10.1177/1556264619850736.

[33] franzke, Bechmann, Zimmer, Ess, the Association of Internet Researchers, *Internet Research: Ethical Guidelines 3.0*.

the Association of Internet Researchers' list of ethical issues[34] and its more recent "Ethical Guidelines 3.0"[35]. These guidelines show the importance for the academic community to deal with digital research practices and to take into account ethical considerations. In relation to those complex issues, while the Association of Internet Researchers (AoIR) advocates a "case-based" perspective rather than a prescription of procedure[36], also points the importance of making explicit contextual interpretations of privacy[37] and data anonymization. Indeed, a fundamental question is the "potential damage" that the users, to whom the data refer, could suffer due to the (re)presentations of the research results[38]. Buchanan and Zimmer[39] emphasize that anonymity and privacy of the observed subjects must be adequately protected where informed consent is very difficult or impossible to obtain[40].

For instance, a social media influencer might even appreciate being the (un)subject of academic studies, but if such influencer regretted having posted images or videos, he/she would want to delete the original posts from their (public) accounts. It is clearly more difficult to remove such material from books and scientific journals.

As Zimmer argued, the condition of Internet data is not analogous to the data we might collect by observing people in physical public spaces[41]. Users may post sensitive information online such as relationship statuses or political opinions, without realizing nor the potential implications of sharing this content[42]. Starting from these data, network analysis and machine learning techniques can allow researchers to generate inferences on some information relating to the users/research subjects compared to what would be possible on the basis of observations in a physical space.

An example of good practices in this sense is provided by the study carried out by Semenzin and Bainotti[43], about Telegram chats in which photographic material of an intimate and non-consensual nature is shared. Although made available by the affordances of the platform, the researchers realized that the access to this large amount of data which, by their nature, constitute a violation of the privacy of the subjects portrayed,

[34] A. Markham, E. Buchanan, *Ethical Decision-Making and Internet Research: Version 2.0. Recommendations from the AoIR Ethics Working Committee*, 2012. Available at: aoir. org/reports/ethics2. pdf.

[35] a.s. franzke *et al*., *Internet Research: Ethical Guidelines 3.0*, 2020, 1-82. Accessed July 13, 2023, https://aoir.org/reports/ethics3.pdf.

[36] A. Markham, E. Buchanan, *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*, 2012, 1-18. Accessed July 13, 2023, https://aoir.org/reports/ethics2.pdf.

[37] The matter of privacy concerning with social networking services represents a subset of data privacy, involving the right of mandating personal privacy concerning storing, re-purposing, provision to third parties, and displaying of information pertaining to oneself via the Internet. Social network security and privacy issues derive from the large amounts of information these sites process each day. Features that invite users to participate in messages, invitations, photos, open platform applications and other applications sometimes constitute the venues for others to gain access to a user's private information.

[38] Highfield, Leaver, *A Methodology for Mapping Instagram Hashtags*.

[39] M. Zimmer, E. Buchanan, Internet Research Ethics", *Stanford Encyclopedia of Philosophy*, 2016. https://plato.stanford.edu/entries/ethics-internet-research. Accessed July 14, 2023.

[40] M. Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton, NJ: Princeton University Press, 2018.

[41] M. Zimmer, "But the Data Is Already Public': On the Ethics of Research in Facebook", *Ethics and Information Technology*, 12 (2010): 313-325.

[42] K. Crawford, M. Finn, "The Limits of Crisis Data: Analytical and Ethical Challenges of Using Social and Mobile Data to Understand Disasters", *GeoJournal*, 80 (2015): 491-502. https://doi.org/10.1007/s10708-014-9597-z.

[43] S. Semenzin, L. Bainotti, "The Use of Telegram for Non-Consensual Dissemination of Intimate Images: Gendered Affordances and the Construction of Masculinities", *Social Media + Society*, 6, 4 (2020). https://doi.org/10.1177/2056305120984453.

could not be simply published but had to be *handled with care*. Thus, the researchers removed any personal or individual references, in order to preserve the anonymity of the users involved. Digital traces represent materials rich in information and important empirical evidence, but their use requires an ethical case-by-case approach (without leading to relativism, but maintaining an approach that considers the evaluation of ethical guidelines and standards shared by the scientific community).

3. CHALLENGES AND OPPORTUNITIES OF POST-API RESEARCH

Until a few years ago, the process of hoarding online content was relatively straightforward using APIs. In the midst of what she calls the "Data Golden Age"[44], scholars were exploiting bugs in the platforms and code and violating terms of service to collect far more data than was technically permitted, while ignoring the terms established by these mega corporations. While scholars are not obliged to protect the business model of tech companies, it is instead our ethical responsibility to avoid putting data first, while neglecting users and their subjectivities inscribed in this data.

Until about ten years ago, we could find several studies that collected and analyzed data extracted from digital platforms without having explicitly obtained the users' informed consent.[45] Similarly, some apps hooked to the Facebook API (e.g. Netvizz: https://up2.fr/Main/Netvizz) have allowed researchers to collect data from users' friends without their knowledge[46].

When the Facebook-Cambridge Analytica (here and after, CA) scandal hit the headlines in March 2018, it shed light on the uses and misuses of personal data by tech companies. Such data are capable of revealing opinions, tastes, sexual or political orientations, the state of physical and mental health, and other sensitive personal characteristics of end users, who have therefore become (*interpassive*) *data-subjects*[47]. CA opened up the case for a very severe privacy issue in the digital world and the expected consequence was the stop of many social platforms on free access to their data[48].

It is therefore not surprising that researchers of digital social life have reacted to this new "post-API era"[49], with research strategies to tap into data that are no longer publicly available. However, the basic relationship between researchers, platforms and digital data seems to have remained substantially the same: platforms and their APIs have always been proprietary "black boxes" of these tech companies, never really intended for academic purposes, but rather for commercial purposes. For instance, Meta

[44] R. Tromble, "Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age", *Social Media + Society*, 7, 1 (2021). https://doi.org/10.1177/2056305121988929.

[45] S. Catanese *et al.*, "Crawling Facebook for Social Network Analysis Purposes", in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, New York, NY, USA: Association for Computing Machinery, 52 (2011): 1-8. https://doi.org/10.1145/1988688.1988749.

[46] This "friends of friends" feature was built into the Facebook API at the time, meaning that scholars weren't then per se violating any of Facebook's terms of service. However, such compliance did not fully absolve researchers of ethical responsibility in regard to those whose data was mined, analyzed, and in many cases, shared with others.

[47] E. Ruppert, "Population Objects: Interpassive Subjects", *Sociology*, 45, 2 (2011): 218-233.

[48] S.M. Özkula, P.J. Reilly, J. Hayes, "Easy Data, Same Old Platforms? A Systematic Review of Digital Activism Methodologies", *Information, Communication & Society*, 2022, 3918. DOI: 10.1080/1369118X.2021.2013918.

[49] Freelon, "Computational Research in the Post-API Age".

today limits access to the API also for research purposes while allowing data sourcing when commercial monetization is more easily obtained.

Both Meta and Twitter made their APIs still partly open to academics. For instance, through CrowdTangle, Meta/Facebook provide a tool that tracks interactions on public content from Facebook pages and groups, verified profiles, public Instagram accounts[50]. Twitter did something similar by opening up for the moment[51] its tweet archive to academic researchers. This is undoubtedly a great achievement, but we must consider that first Twitter is little used in Italy[52], and to answer questions about the daily micro-stories t shared on these social media (which was the object of the research that we will present), it is necessary to experiment with an observation method.

Recent research conducted on a sample of Italian researchers[53] highlight how limits on social data access seem to have not really created a "post-API" scenario, but it is turning research practices upside down, with mixed implications: while researchers are fruitfully experimenting with innovative approaches, there is a chance for a sort of "migration" to the few platforms that freely grant their APIs, with critical consequences for the quality of research. Furthermore, the cancellation of APIs may have compromised academic work in progress[54]. The conditions of producing digital research have worsened, given that empirical inquiries are increasingly oriented to "easy-data" environments such as Twitter[55].

After 2018, digital researchers sought to take a more critical look at how the academic community collected and analyzed data when it still seemed so abundant. They then incorporated those reflections to inform future analysis of big data so that greater respect for the rigor, ethics, and broader social values one should expect in not-for-profit research was given[56].

All in all, to consider the entanglement between subject and data means following ethical guidelines that limit the use of online content that researchers can still access via APIs today, since they are considered public, or rather, publicly available: such as those on forums, blogs, pages and public profiles of Facebook and Instagram, or on Twitter.

Doing Internet research in the post-API era means, on the one hand, using methodological imagination and developing strategies to 'repurpose' digital methods in a post-API research environment[57]. On the other, when data of/on the subject/user is

[50] N. Shiffman, "Crowd Tangle for Academics and Researchers", Help.Crowdtangle, https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers.

[51] As we write this paper, Twitter (recently bought by Elon Musk) announced the shutting down of its API in February 2023, with therefore imminent limitations also for the academic world.

[52] While Facebook and Instagram are utilized by 78% and 73% respectively, Twitter is only used by the 26% (Wearesocial, 2023).

[53] Trezza, "To Scrape or Not to Scrape, This Is Dilemma".

[54] J. Hemsley, "Social Media Giants Are Restricting Research Vital to Journalism", *Columbia Journalism Review*, 2019. Retrieved February 13, 2023, from: https://www.cjr.org/tow_center/facebook-twitter-api-restrictions.php?fbclid=IwAR1uULkcGqcOrQYSagYkfSaKciKGK5t2x_Q5hnoOd38CGs02ND_oVULdpns.

[55] Twitter seems today the most (over-)studied social media, because it offers relatively open data access. Its public Search API allows researchers to gather tweets posted up to 7 days earlier, while the public Streaming API permits capture of tweets in real time. Twitter now requires scholars to undergo review for API access, and the company only allows each researcher to use only one app to query the APIs. However, despite those limitations, APIs still allow academics to gather large amounts of Twitter data, no matter their financial resources.

[56] Tromble, *Where Have All the Data Gone?*, 2021.

[57] A. Caliandro, "Repurposing Digital Methods in a Post-API Research Environment: Methodological and ethical implications", *Italian Sociological Review*, 11, 4S (2021): 225-225.

collected – it requires reflecting on the various phases of research so that ethically responsible choices can be consciously taken relating to: privacy, anonymity and informed consent.

4. THE "7 FRIEND FOR 7 DAYS" STUDY

The study analysed the content created by users on social media platforms, specifically focusing on the types and subjects of micro-narratives, as either Stories or Posts on Instagram or Facebook. This paper will reflect on the ethical implications of the research design and of the results (published elsewhere)[58].

4.1. *The method*

The empirical project on which this paper draws was conducted in 2019 with the goal to investigate the production of users' contents (*unstructured data*) on their personal Instagram and Facebook profiles. Compared to previously published studies that aimed to analyze the different usage practices and the underlying motivations[59], our inquiry opted to analyze the level of discourse, i.e. what is published by users (in terms of posts and images).

As highlighted earlier, Meta's APIs posed limitations even for academic investigations, precluding the possibility of downloading the contents that users publish on private Instagram and Facebook profiles.

Our methodological proposal needed to be therefore configured in the context of post-API studies, indeed, it would not be "technically" possible to access these data, since they were embedded in personal user profiles. Therefore, the data collection phase was not based on the automatic extraction of contents from social media[60], but contents were considered as data collected through observational methods[61].

The project design was realized through a manually coded sample of users' generated contents (textual posts or photo + related captions), so that the analysis could be carried out on the coded data, for which the original contents have been completely anonymized.

The research process took place as follows: researchers asked a team of collaborators to select 7 "reciprocal" followers on the Instagram platform and 7 friends on Facebook, and then to observe the contents they published on their profiles. The coders consisted of 65 students who participated in a workshop on "Digital Methods", scheduled for the last year of the "Communication, Media and Advertising" Master degree course at the IULM University of Milan. They followed several extensive training phases, both to make them understand the research method and to pedagogically stimulate their critical approach. They independently coded 50 UGCs with the purpose of measuring the inter-coder agreement levels. The percentage of agreement was 86.5%-100%. Krippen-

[58] E. Risi, Gandini A., eds., *Tracce digitali e ricerca sociologica: Riflessioni ed esperienze di sociologia digitale*, Milano: FrancoAngeli, 2023.

[59] B. Kim, Y. Kim, "Facebook versus Instagram: How Perceived Gratifications and Technological Attributes Are Related to the Change in Social Media Us-Age", *The Social Science Journal*, 56, 2 (2019): 156-167.

[60] Bainotti, Caliandro, Gandini, "From Archive Cultures to Ephemeral Content, and Back: Studying Instagram Stories with Digital Methods".

[61] B. Smart, K. Peggs, J. Burridge, *Observation Methods: Four Volume Set*, Sage Publications Ltd, 2013.

dorff's alpha range for the analysis sheet's 35 variables was 0.712 – 1.000, which are considered of adequate reliability (Krippendorff, 2004). The analysts were fully awareness and consent for their work to be incorporated in scientific papers.

Regarding Instagram, the analysts monitored the profile every day, for an entire week, taking note on a content analysis sheet (uploaded on SurveyMonkey platform) about some characteristics of the contents posted by the users/subjects observed (for example, the type of subject of the photos, the topic of the post, the length of the caption, the presence of emoticons, tags, hashtags and so forth). On Facebook, the coders proceeded with a similar approach: every day, for seven days, all the posts published on the personal profiles of the followed users were observed and coded through a content analysis. Stories were not included on this social media.

As a result, the data (n = 3,672 stories and 240 posts; and a total of 1,091 photos) were analyzed between April 6 and May 12, 2019. The UGCs refer to a sample of subjects with an average age of 23 for Instagram (mostly students -79%- with only 21% workers) and 27 years for Facebook (again mostly students, 73%).

These aspects are interesting for two reasons. First of all, because we focused on how stories are constructed on social media through visual-narrative elements, and what contents are shared by a sample of young people. Second, data were coded by "young researchers", who were trained both on how to implement the method and on the important ethical implications that were involved in that methodology. The project was also part of a process aiming to raise critical awareness about the use of digital platforms, indeed, the activity was explained to the coders as an exercise in which observers had to critically reflect on the daily practices connected to the use of social media (inspiring by Markham's critical pedagogy approach).

4.2. *Ethical issues*

As the research developed, we began to think about how to analyze and present the data, trying to follow an ethically responsible approach to the subjects inscribed in that data. We will outline below the challenges we encountered and how we tried to solve them. The "ethically important moments"[62] we experienced in our project emerged from three main aspects, which could be considered as potentially controversial points in the collection and presentation of results.

To address these issues, we have adopted a number of strategies, in an iterative process of adjustment and redefinition, which occurs whenever research opportunities are challenged by ethical issues.

The first of these ethically important moments concerns the generally intimate and potentially sensitive nature of the posts and above all of the photos published on the 90 accounts (between Instagram and Facebook), which are, to all intents and purposes, personal profiles. Considering that the collected UGCs focused on moments of users' daily lives, i.e. involving private relationships and experiences, we had to recognize and share with the analysts the responsibilities we carry when working with these digital data, particularly to the people represented in those data. According to Zook *et al.*[63], it

[62] M. Guillemin, L. Gillam, "Ethics, Reflexivity, and 'Ethically Important Moments' in Research", *Qualitative Inquiry*, 10, 2 (2004): 261-280. https://doi.org/10.1177/1077800403262360.

[63] M. Zook, S. Barocas, d. boyd, K. Crawford, E. Keller, S.P. Gangadharah, "Ten Simple Rules for Responsible Big Data Research", *PLoS Computational Biology*, 13, 3 (2017), e1005399. doi:10.1371/journal.pcbi.1005399.

is not enough to reduce the issue in a binary way between public or private data: that is because when social media data are used, the concept of privacy becomes "contextual and situational". The choice was to analyze the data through codified categories and to do it in an aggregate form. In this way, it would not have been possible to recover the original profile of the users or trace them, just as it was not possible to trace the interviewee in a traditional survey[64].

So, based on our approach and its self-limitations, we were only able to engage with the data at the aggregated level. The decision to consider them as small data, reflected the same care for ensuring the participants' anonymity in traditional research. This sometimes requires omitting revealing details despite their analytical value.

The second ethical moment consisted in the exclusion from using the content of users who had not granted permission to be observed. We consider that social media contents address audiences made up of different actors; for ethical issues it is essential to consider whom the posts/photos are actually addressed to.

Informed consent is nearly never a possibility and isn't easy to guarantee when the examined posts appear to be "publicly available", especially if extracted in "large quantities" from social media monitoring platforms or through the platforms' APIs[65]. While deploying "contextual integrity" tactics and evaluating the scale of archived social media materials represent significant ethical improvements[66], in the case of small data analysis, getting in contact with creators of digital materials ties to a fundamental ethical dimension for Internet researchers[67].

We therefore chose to engage with users: each analyst was urged to ask for informed consent (before proceeding with the observation) in analyzing and representing the observed UGCs. Accordingly, the analysts (coders) contacted through Instagram "direct-message" function each subject, informing the users about the aforementioned elements and waiting - before proceeding with the observation and coding – for them to agree to the collection by declaring their informed consent.

In terms of the entanglement between subjects and data, the request for informed consent entails the fact that the observed users/subjects (or rather, whose published contents are analyzed) are made explicit: the purpose of the research, the methods used, the possible research results, and the possible risks that the disclosure of these results may bring.

We could also have asked the analysts to get permission to use the posts or photos of the analyzed users. This would have been done in order to be able to reproduce these contents as research output, after anonymizing names and references to users in order to make posts and images less easily identifiable and searchable. But we didn't do it, instead limiting ourselves to the exposure of only quantitative data based on content analysis.

This represented the third ethically important moment, namely about the reproduction of some UGCs in presentations or academic papers. Especially because most of the data collected refer to ephemeral contents (Instagram Stories), and therefore it is assumed that the user, in using these social media affordances, really wants them to no

[64] G. Di Fraia, E. Risi, *Empiria. Metodi e tecniche della ricerca sociale*, Milano: Hoepli, 2019.

[65] There are digital platforms that exploit the APIs of websites and social media to collect and aggregate a series of data, also offering various options for viewing the data itself and a certain number of indicators.

[66] S. Lomborg, "Ethical Considerations for Web Archives and Web History Research", in *SAGE Handbook of Web History*, edited by N. Brügger, I. Milligan, London: Sage Publication, 2018, 199-219.

[67] K. Eichhorn, *The End of Forgetting: Growing Up with Social Media*, Cambridge, Mass: Harvard University Press, 2019.

longer be viewable after 24 hours. This reflexivity made us choose to treat the data in an exclusively quantitative and aggregated way, thus avoiding the insertion of examples and quotations, even if they were interesting.

We preferred to be cautious in the face of the ethical challenge regarding the potentially sensitive nature of posts and photos (which in some cases portrayed couple and family relationships), not only in the present moment, but with respect to its repercussions in the future.

The request form for informed consent (sent – as mentioned – before observing and analyzing the UGCs of social media profiles), explained that the data obtained were going to be treated in such a way as to respect the privacy of users: e.g. without the disclosure of the results, given the potential harm of their publication. Moreover, the form specified that the data were going to be collected only for academic research purposes. In other words, we made sure that the dataset obtained was not going to be shared with third parties[68], and none of the posts or photographs would be mentioned or illustrated in scholarly publications.

5. Concluding remarks

In this paper we cover an important topic, namely research using Internet platform data and the ethical issues surrounding the usage of such data. We proposed the approach exemplified by "7 friends for 7 days", as a case study to investigate the specific challenges of working with Instagram and Facebook data within the context of research data access challenges.

As we have seen, the research material is often based on the contents that users publish online (constituting unstructured data), which inevitably intercept some intertwined ethical dimensions: the statute of data (whether public or private); the awareness that users have of the possible use of their data; and the question of privacy related to data anonymization and informed consent.

While social research in digital societies requires rigor in ethical terms, particularly in the access and management of personal data, the research process and the iteration of procedures in this context have been consolidated through attempts sometimes experimental and creative in nature, which have pushed researchers to go 'beyond' the given prescriptions, and which points how Internet research is an active and evolving field, which also follows the evolution of digital platforms.

While ethics-aware approaches may seem limiting, both in the case of research on contents that are still "technically" publicly available, in our view it is important that scholars carefully consider those ethical aspects involved in handling user-generated content.

Our argument favors a complex understanding of data (whether big or small) in a social media environment as human-shaped artifacts, which we suggest calls for a consciously ethics of care, taking up the commitments to the subjects/objects of study imbricated in data.

Social media platforms are a unique and incredibly fruitful resource for studying daily life but there are ethical concerns around a researcher's approach to this data.

[68] C. Fiesler, N. Proferes, "Participant' Perceptions of Twitter Research Ethics", *Social Media+Society*, 4, 1 (2018). DOI: 2056305118763366.

It is imperative that we consider whose stories are being told, who is equipped to tell them, and what kinds of vulnerability and harm we might encounter and bring up when doing so[69].

As we said, another important challenge in Internet research is grappling with the distance between researchers and the researched: as Luka and Millette pointed out[70], a feminist-based ethics of care, should be adopted at every stage of the research, especially when the data come from people's lives such as stories, opinions, and private images. The researcher should in fact handle with care also those contents that the user publishes on his/her profile, thus un-reflectively making them "public" to all, due to lack of attention or awareness in managing the privacy levels of the platform.

Thus, it is important that we raise critical awareness how to do research taking into account these ethical principles, in self-reflexive process that guide the dialectical and reciprocal relationship between observing subject and observed objects/subjects. For instance, in the case of exploring users generated content, we invited the observers to re-center human and move towards an ethics of care[71] for engaging with the observed data. It was also a way of reciprocally observing others through ourselves, taking care of the small "portions" of life that others allow us to observe.

Admittedly, the reflections put forth in this article have some relevant limitations. First, the UGCs analyzed refer to a number of observed subjects close to a convenience sample and therefore it doesn't allow for generalizability. Second, the points discussed here stem from only from a case study which, furthermore, was conducted by the authors themselves. Despite all that, we think this article can start a fruitful conversation on some important ethical issues in social media research, rather than proposing generalizable results or universal theoretical concepts.

[69] franzke *et al*., *Internet Research: Ethical Guidelines 3.0.*

[70] Luka, Millette, *(Re) Framing Big Data.*

[71] M.E. Luka, M. Millette, J. Wallace, "A Feminist Perspective on Ethical Digital Methods", in *Internet Research Ethics for the Social Age: New Cases and Challenges*, edited by M. Zimmer and K. Kinder Kurlanda, Bern, Switzerland: Peter Lang, 2017, 21-38.