Gerwin van Schie; Irene Westra; Mirko Tobias Schäfer

Get Your Hands Dirty: Emerging Data Practices as Challenge for Research Integrity

2017

https://doi.org/10.25969/mediarep/12434

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
van Schie, Gerwin; Westra, Irene; Schäfer, Mirko Tobias: Get Your Hands Dirty: Emerging Data Practices as Challenge for Research Integrity. In: Mirko Tobias Schäfer, Karin van Es (Hg.): *The Datafied Society: Studying Culture through Data*. Amsterdam: Amsterdam University Press 2017, S. 183–200. DOI: https://doi.org/10.25969/mediarep/12434.

13. Get Your Hands Dirty

Emerging Data Practices as Challenge for Research Integrity

Gerwin van Schie, Irene Westra & Mirko Tobias Schäfer

Introduction

In November 2014 two interns (the first two authors of this chapter listed above) at the Utrecht Data School started investigating an online discussion forum for patients under the supervision of Mirko Tobias Schäfer (this essay's third author). Without his knowledge and without any prior knowledge of scraping websites, the two students downloaded 150,000 patient profiles (which included, amongst other information, age, location, diagnoses and treatments related to these patients), using a (90-euro) off-the-shelf scraper tool[1], without informing these patients or requesting consent from them or the platform providers. The plan to first explore the data (taking the necessary precautions to keep the data confidential) and later, after formulating a research question and hypothesis, to ask permission to conduct in-depth analysis of data relevant for our research, was never realized. After a few days of acting like 'information flâneurs' (Dörk et al. 2011), browsing through the data without specific questions or goals in mind, we were notified that our department's supervisors had terminated the project due to concerns about research ethics.[2] Their decision prompted us to rethink our actions and to question our research practices as well as existing research standards. Assuming that the rather novel data sources and practices of analysis were disrupting the traditional research process and contradicting established guidelines in research ethics, we found that these events provided the inspiration to revisit research ethics concerning big data research.

1 Outwit Hub, www.outwit.com/products/hub/.

2 After learning about the data scraping, the project supervisor (Mirko Tobias Schäfer) immediately reported the project to the director of the research school who informed the vice dean of research and the ethics committee. While the decision was pending all data were stored in a secured environment and access was limited to the investigators and documented accordingly. After the board's decision to terminate the project the data were securely deleted. It must be emphasized that the students' activities – despite being disputable – were considered legal as the information was openly available.

Although the forum we investigated was not technically a social network site (SNS), we think that the issues we will discuss in this chapter are very similar to the ethical issues relevant to the investigation of SNSs. The characteristics of available (big) data sets and emerging data practices do not always afford a practice that complies with traditional standards of research integrity. Such standards were very much informed by events marked by severe human rights violations and scholarly misconduct. They responded to incidents in which the lives of 'human subjects' were harmed. Although current research practices do not necessarily cause physical pain, they may violate personal integrity and fail to meet privacy standards by accidentally revealing someone's personal identity or sensitive information about individuals who are part of the sample. When investigating a Web forum for patients afflicted with a specific disease, the authors of this chapter experienced the various promises and pitfalls of digital methods. Drawing from this experience, we reviewed existing standards of scholarly research practice, focusing particularly on media studies. This chapter revisits the formative guidelines that provided the historical basis for current ethical research guidelines, including the Nuremberg Code (1947), the Declaration of Helsinki (World Medical Association 2013) and the Universal Declaration on Bioethics and Human Rights (UNESCO 2006). We will argue that the existing ethical guidelines are relics of discourses and eras that have very little to do with Big data research as it is now conducted. In addition, referring to a case study that describes our own experiences, we will explain how big data research on social networking sites makes the concept of informed consent, a basic principle of all current guidelines, practically infeasible. Building on the guidelines that have been written for internet researchers (Markham & Buchanan 2012), we will conclude with a proposal for a research structure consisting of three stages, each with its own ethical considerations: design, safe data exploration and data analysis.

Big Data and the Humanities

Under the label 'digital humanities', several novel research practices have been developed within social research and media and cultural studies (e.g. Berry 2012; Burdick et al. 2012). For a long time these domains have been strongholds of qualitative research, participatory observation and hermeneutic approaches to textual analysis. Now, new data-driven and computer-aided methods have stirred up dust within departments that had seldom been compelled to question their professional standards of

conduct. The rapidly changing situation is now marked by unprecedented access to vast data resources and innovative tools to collect and connect large numbers of data points. As a result, some of these digital humanities projects require skilful interdisciplinary cooperation. Researchers even seek out collaborations with programmers, entrepreneurs, corporations and organizations who contribute technology support, data collection, data hosting or other services. On the other hand, as in our case study, there are now tools that allow researchers with relatively limited technical skills to adopt some of the new practices. Additionally, researchers and their academic institutions have become concerned with the data samples and the practices of investigating the data (Rieder & Röhle 2012). The so-called T3 study (Lewis et al. 2008) and the more recent Facebook study (Bond et al. 2012) have come to the attention of institutional review boards and scholars alike, who point to the need to consider privacy concerns and informed consent when using (big) data from social media platforms (Zimmer 2010).

However, scholars cannot neglect the unprecedented access to new data resources. Historians, literature scholars and information and library scientists quickly recognized that digitized texts provide rich data with which to address novel research questions. Within media studies, the added value of 'natively digital' elements was quickly recognized and used for research (Rogers 2009). Pioneered by media scholar Richard Rogers, it led to the emergence of a set of practices and tools to systematically collect and analyse these data from Web platforms (Rogers 2013). Using digitized cultural artefacts from films to graphic novels to the metadata of Instagram photos, Lev Manovich (2012) applied his approach of 'Cultural Analytics' to use analysis software to detect patterns of cultural production and media use. These new practices are rapidly changing the field of media studies in general and new media studies in particular, as knowledge of these tools and practices is increasingly a requirement in academic hiring. In the field of sociology, the emergence of newly accessible data sources and novel analysis tools has led to a debate that revisits the notion of the 'empirical'. It became clear that the sheer size and variety of the data problematize 'stock-in-trade analytic methods' (Abbott 2000: 298) and that the existing methods of explanation were not suited to the 'increasing availability of a wide range of data that previously was not easily accessible, but is now routinely collected as part of information and communication techniques' (Adkins & Lury 2009: 15).

In this article, we consider how these recent changes in tools and practices affect research methods and ethics. Informed consent – the principle value of research ethics in the social sciences – is under pressure due to two

technological advances: the rise of the internet and big data technologies. In the following paragraphs we will explain the origins of informed consent and the way it should be dealt with in the field of big data research on SNSs.

Finding Suitable Guidelines

As early as 2002, Michelle White discussed the limitations of the use of one single guideline to govern the spectrum of possible ways of conducting research on the internet:

> It seems unlikely that any single guideline for Internet research ethics can resolve conflicts between the disciplines. For instance, the 'Protection of Human Subjects' document requires that 'risks to subjects are reasonable in relation to anticipated benefits, if any, to subjects, and the importance of the knowledge that may reasonably be expected to result' and MLA mandates that 'whether a line of inquiry is ultimately useful to society, colleagues, or students should not be used to limit the freedom of the scholar pursuing it' (Code of Federal Regulations. 2001, Title 45, Part 46). Obviously, a more careful articulation of both 'subject' and 'representation' would aid in these considerations. At the moment, guidelines for Internet research have not addressed such disciplinary conflicts and have instead almost completely ignored the conventions in a number of Humanities disciplines. (White 2002: 255-256)

As an emerging digital humanities discipline, big data research has exactly these problems. First, because it takes a hypothesis-generating approach to data, often the usefulness of a line of inquiry is not known beforehand. On an institutional level this is problematic, since requesting funds or grants for research often requires that research objectives be described in advance. In other instances, ethical guidelines can complicate the application, as many issues are ambiguous and unclear, such as the extent to which public profiles and publicly posted information are subject to privacy regulations. Other cases have sparked criticism after publication for the supposed disregard of ethical guidelines. Frequently mentioned in this regard is the so-called Facebook study, which received wide media coverage (Bond et al. 2012). A massive outcry about the researchers' supposedly reckless behaviour arose when it was revealed that user timelines were being manipulated to investigate the emotional impact of Facebook's news feed on users (Puschmann & Bozdag 2014; Schroeder 2014). Informed consent

was construed as being given through the user's acceptance of Facebook's terms of use at sign-up. This act was conveniently interpreted to signify the user's agreement to their data being used for research. The data for the study was generated by manipulating the timelines of a large group of Facebook users, in total 60,055,176 profiles. Facebook employees anonymized the data before handing it over to the researchers. Because the researchers were not dealing with data that could be connected to identifiable individuals, they did not classify the research as human subject research and assumed they did not have to comply with regulations regarding such practices (Carberry 2014, in a press release by Cornell University Media Relations). In countries not making use of IRBs, the solution to the problem of possible ethical breaches has to be sought in more general guidelines governing the conduct of individual researchers. Michelle White makes a good start by proposing the use of ethical principles stemming from other disciplines. Regarding informed consent in internet research, she offers an argument based on the relation and difference between human subjects and their online representations in the form of profiles and accounts (2002: 249).

Informed Consent

Informed consent has been an integral part of all guidelines concerning human subject research in the medical, sociological and psychological fields. How this principle should be used in the field of internet research on SNSs is still heavily debated. Psychologist Ilka Gleibs (2014) has discussed ethics in large-scale online studies on social network sites. She argues that informed consent of participants is needed when one wants to use data from these sites:

> The use of informed consent is important because it allows participants to make a choice and signals their willing participation. As researchers we show respect for the individuals' autonomy, which is a fundamental ethical principle. (Gleibs 2014: 5)

Referring to the controversial T3 study (Lewis et al. 2008), Michael Zimmer (2010) emphasizes the need to hold on to existing research standards, arguing that one cannot be ethically lax simply because these data are freely available via Facebook. The recent controversy about the Facebook study (Bond et al. 2012) mentioned above, which manipulated Facebook timelines without users' consent, indicates that certain research practices conflict

with the traditional understanding of research integrity. Indeed, for many SNS research projects, informed consent represents the underlying pact between researcher and the subjects in the 'field':

> [I]n order to represent and analyse pertinent social phenomena, some researchers collect data from social media without considering that the lack of informed consent would in any other form of research (think of psychological or medical research) constitute a major breach of research ethics. (Zwitter 2014: 5)

To understand the conflicting visions of how to investigate social phenomena on Web platforms, we recall how ethical standards for research including human subjects came into being. The Nuremberg Code, the Declaration of Helsinki and the Universal Declaration on Bioethics and Human Rights are three regulatory guidelines that are often cited in academic discourse on human subject research (White 2002; Buchanan & Ess 2008; Markham & Buchanan 2012; Gleibs 2014; Dumas et al. 2014: 375). The Nuremberg Code[3] is one of the first documents on human rights that characterizes voluntary informed consent as a fundamental ethical principle (Grodin 1994). But one of its problems, according to physician researchers, is that it did not take clinical research on children, patients or mentally impaired persons into account (Annas 1992: 122) The Declaration of Helsinki can be seen as a more elaborate and more easily applicable document than the Nuremberg Code (*ibid.*). One big difference concerns the expertise of the writers who wrote the documents: the Nuremberg Code was issued by judges (who adopted and expanded ethical principles initially provided by psychiatrist and neurologist Leo Alexander), whereas the Declaration of Helsinki was written by physicians. Another difference is that the latter has been revised regularly: six revisions have been made since the first version appeared in 1964. After all these years, the Declaration is even referred to as 'the most widely accepted guidance worldwide on medical research involving human subjects' (Christie 2000: 913). Ethical guidelines should not be static, and the Declaration of Helsinki proves to be a good model of a set of protocols that has been adapted to meet evolving needs and situations.[4] The Universal

3 'Trials of War Criminals before the Nuremberg Military Tribunals Under Control Council Law 10' (Washington, D.C.: Superintendent of Documents, United States Government Print Office, 1950). Military Tribunal 1, Case 1, United States v. Karl Brandt et al., October 1946 – April 1949, Vol. I, pp. 1-1004; Vol. II, pp. 1-352 (1949).

4 For more substantive information, the article 'The Revision of the Declaration: the past, present and the future' by Robert V. Carlson, Kenneth M. Boyd and David J. Webb (2004), is recommended.

Declaration on Bioethics and Human Rights is the first document that binds UNESCO member states – 195 countries – to one declaration (Berlinguer & De Castro 2003). As its title indicates, its purpose is to provide guidelines for ethical issues 'related to medicine, life sciences and associated technologies as applied to human beings, taking into account their social, legal and environmental dimensions' (UNESCO 2006).

Ethical Decision-making in Internet Research

The Nuremberg Code, the Declaration of Helsinki and the Universal Declaration on Bioethics and Human Rights were all binding (to varying degrees), but each was written under different circumstances, employed different discourses, and was conceived with different kinds of research in mind. According to Dumas et al. (2014: 375) there are, in general, two features evident in most research regulations around the world: first, regulations are often written in reaction to unacceptable research practices (as with, for example, the origins of the Nuremberg Code, which was formulated in the wake of Nazi atrocities); second, these regulations often do not take into account evolving forms of technology (such as possibilities for data gathering). Writers who have accounted for the current state of technology, such as Zwitter (2014: 375) and boyd & Crawford (2012), have been vague. The closest attempt to a set of guidelines for internet research has been written by the Association of Internet Researchers (AoIR), an academic association focused on the cross-disciplinary field of internet studies. This association promotes critical and scholarly internet research. It drafted a first version of the AoIR Ethical Decision Making document in 2002. A second version appeared in 2012, as the association had decided that a revision was in order because the scope and context of internet research had changed rapidly. The AoIR encourages internet research independent of traditional academic borders (AoIR 2015); its basic ethical principles rely on, amongst others, the Nuremberg Code and the Declaration of Helsinki: 'We accept them as basic to any research endeavour' (Markham & Buchanan 2012: 4). The problem of internet research that has to be faced, according to AoIR, is caused by the dynamic evolution of the field of research:

> This dynamism is reflected in the fact that as of the time of this writing, no official guidance or 'answers' regarding internet research ethics have been adopted at any national or international level. (*ibid.*: 2)

The association has no intention of providing 'definitive' regulations that would foreclose further discussion about how to do internet research in an appropriate way: 'We emphasize that no set of guidelines or rules is static; the fields of internet research are dynamic and heterogeneous' (*ibid.*: 2). Thus the Ethical Decision Making document – and this is ultimately a shortcoming – proposes an extensive list of 'Internet Specific Ethical Questions' to 'prompt reflection about ethical decision making within the specific confines of one's study' (*ibid.*: 8): it is up to the researcher to determine which questions are relevant for the research being conducted, and which ones are not. The field of big data has not yet been discussed extensively; however, in earlier work, Markham has discussed certain characteristics of qualitative research ethics that have interesting similarities with the way big data research is being done:

> Ethics is considered an a priori stance, often regulated more than felt by the researcher. Research design is often considered a procedural or logistic matter, mostly followed, not questioned, particularly if the researcher is within junior ranks of the profession or working within a discipline that values adherence to particular approaches. The consideration of research design as a given is founded in epistemologies that value precision, replicability, validity, and objectivity, all of which require a priori determination of activities. Any interference in the procedures or disruption of pre-determined standards is discouraged because it may invalidate the study. This is antithetical to the idea of context sensitivity and reflexivity. (Markham 2006: 43)

Several problems need to be addressed to ensure that in the future, researchers focusing on big social data can do their research in an ethical way. Markham & Buchanan (2015) notice the different fundamental values expressed in the European and American guidelines. Whereas the UNESCO code takes a de-ontological approach (some boundaries should never be crossed), the American Belmont Report has a utilitarian basis (benefits can outweigh downsides). In the following paragraph we will explain how one need not be forced into a choice between a utilitarian and a de-ontological approach if one adopts a stance of ethical pluralism (Ess 2006; 2007). The underlying question here concerns the possibility of research interest trumping research ethics. This matter can entail harsh consequences when one is dealing with, for example, found (or stolen) data sets. One well-known case is the Ashley Madison data leak, in which user profiles on a popular adulterous dating site were made public. The leak of

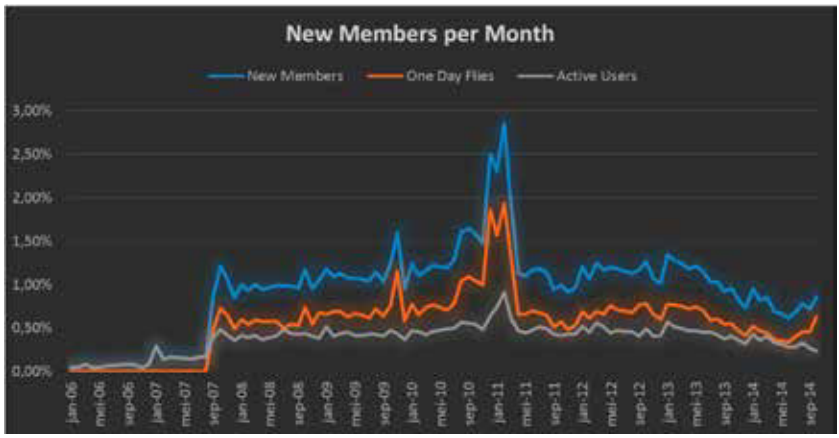
hacked data revealed not only subscribers' personal information but also the site's heavy use of bots so that it would appear to have far more female users than it actually did and to encourage communication on the site (Newitz 2015). Also well known is the so-called Cablegate case. At the end of 2010, WikiLeaks publicized internal communications of the American diplomatic corps (*The Guardian* 2010). As these cables have still not been declassified by the American government, several US-based journals of political science have been declining papers that use the cables as sources of information, effectively preventing scholars from using crucial information for research (Michael 2015).

Another important issue is the possibility of using new tools for data gathering or scraping. With these tools, websites and online communities can be studied even if they would not like to participate in research. Software like Import.io and Outwit Hub make it incredibly easy to scrape databases from public websites and make them searchable and usable for research. Additional tools can be used to anonymize individuals in the data sets. Platform providers prevent automatic scraping through the blocking of suspicious IP addresses, but such measures can be circumscribed through the use of VPNs or proxies. Marketeers, spammers and researchers routinely employ such tools to gather information. Often neither the tools nor the collection of data are illegal, even if the terms of use of a platform state otherwise. Researchers therefore find themselves in a dilemma. Their fair use guidelines and the widely shared imperative of informed consent require them to inform populations on platforms and platform providers about what they are doing. Michael Zimmer emphasizes that the frequent excuse that the data is being made publicly available is unacceptable (2010). However, as we will point out below, it is not always feasible or desirable to comply with the consent requirement.

Case Study: Big Data Research Without Informed Consent

In November 2014, we started to conduct a big data research project on a discussion board of patients afflicted with the same illness. After logging in, the profiles of all members were open for inspection. This included all the information they chose to share with the community. To explore this information we used a scraping tool to scrape all information from all profiles. We found out that about 15 percent of the community had filled in quite detailed information about their specific condition. Similar to what happens quite often in a big data research project, we focused on the

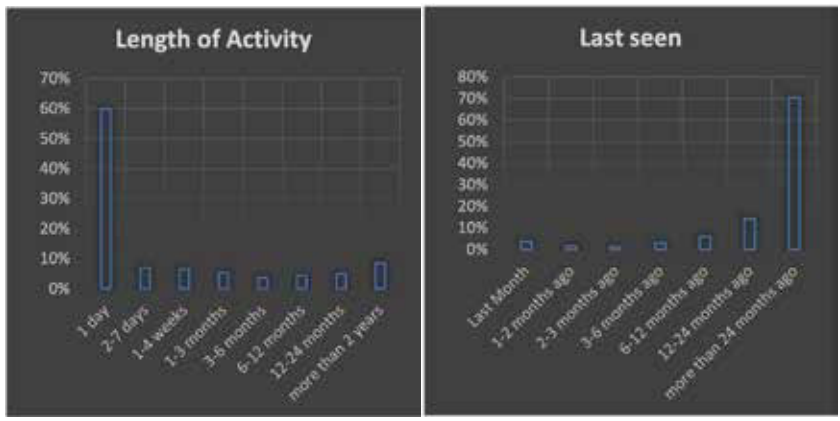
Fig. 13.1: New members of Forum X over time.



information that seemed most valuable at first, only to find ourselves at a dead end after about a week. The medical information could be of value only if we could connect it to specific behaviour on the forum itself. Since we gathered only profiles, and not the conversations in the fora and topics sections of the discussion board, we considered this direction of study to be a dead end. In the following few days, two dates quickly became important in our research: the date the profiles were created and the date a profile was last active. We could measure the forum's growth over time by adding up all the dates of profile creation (see Figure 13.1).

We thought that to understand the function of this forum in this particular community of patients we had to gather qualitative data, too. Therefore, we tried to find a representative sample of people that looked most promising for providing information about their use and media practices: namely, the long-time forum users who were still active. To do so we created two graphs: one with all the profiles sorted by the date of last activity (see Figure 13.2), and one with all the profiles sorted by length of activity (see Figure 13.3). We measured the length of membership activity by subtracting the date of creation from the 'last seen' date. When the groups that were active for more than two years (8 percent of the profiles) and had visited the site within one month before the research began (4 percent of the profiles) were combined, we were left with a sample of 1.2 percent of the total population for further qualitative research. As a by-product we found out that 59 percent of all members had been active only for one day. These 'one-day flies' had either only made an account and never logged in,

Figs. 13.2 and 13.3: Length of activity and 'last seen' date



or had logged in the next day and never returned. In addition, we found that about 70 percent of the profiles had not visited in the last two years. As can be expected, there is a big overlap of almost fifty percent between these groups of people.

The inactive group and the 'one-day flies' make visible the problem of informed consent for this forum. The 70 percent of inactive users would probably never reply to a request for consent, simply because these users were no longer active on the forum. We also expected that a big part of the group of one-day flies represented people who would never respond to a request for informed consent – for two reasons. First, amongst the one-day flies there were a certain number of fake profiles (used, for example, for spamming), since there were a lot of homepage URLs that referred to websites concerned with porn, cosmetics, real estate and other subjects not related to the specific disease or the forum's other conversations. Second, we expected that a large part of the one-day flies also represented inactive users, even if their accounts had been made within the previous two years. In reality, the amount of non-responding accounts, accounting for overlap between the two inactive groups, will be close to 80 percent of the total population. Adding to that, we would like to state that, since we were taking a different direction in our research, these accounts were not of any interest to us and would never have been part of the final sample. Still, existing ethical guidelines would have demanded informed consent from all of these users.

The approach of this case-study can be seen as exemplary for big data research with data generated through user activities on an SNS. As we

showed, it is highly impractical and maybe even impossible to get informed consent from the entire community or network. A second observation is that the use of the term 'human subjects' is debatable in the context of big social data as not all profiles represent actual users. We strongly believe that big data research can be performed in an ethical fashion without getting informed consent from the whole population of an online service. We therefore propose an alternative way of dealing with these subjects in big data research on SNSs.

Proposal for a Three-step Research Process

Drawing from practical experience, we developed a concept for integrating ethical decision-making into the research design process. With reference to Markham and Buchanan, we also argue for guidelines rather than strict codes. It is necessary to adapt the research design to the need for ethical decision-making. The degree of the potential privacy breach or damage that can result from research will have a significant effect on ethical decision-making: when a scholar investigates an SNS, a user forum or an online community, the vulnerability of the target demographic is relevant to the decision-making.

The research that initiated this paper dealt with an online discussion forum for patients afflicted with a certain illness. The website advertised that it had more than 100,000 members. When two interns with limited technical abilities started to investigate the forum we, naively, thought that the ethical framework could be formed simultaneously with the design of the scraping process. We never expected to be able to acquire the complete database in less than two days. Requesting informed consent would have meant that a vast number of inactive profiles would never have been found. Fake, deceased or inactive members are not able to give consent.

Only after discussion with our supervisors did we understand the magnitude of the actions we had undertaken. It was decided that we should immediately terminate the research and destroy all the data we had acquired. The argument that we had only gathered data that was publicly available could not cancel out the fact that we had not asked for consent. We have the strong conviction that researchers should be able to carry out this type of research in the future. A lack of consent from the users of public fora or platform owners or administrators themselves does not have to be an obstacle to an ethically sound research design. To

support this hypothesis, we point to Richards & King (2014), who recognize a difference between privacy and confidentiality: 'With the power of big data to make secondary uses of the private information we share in confidence, restoration of trust in the institutions we share with, rests not only with privacy but in the recognition that shared private information can remain "confidential".' (*ibid.*: 413) To test the boundaries, they advocate experimentation:

> A central part of this experimentation, if we are to have privacy, confidentiality, transparency, and protect identity in a big data economy, must involve informed, principled, and collaborative experimentation with privacy subjects. (*ibid.*: 431)

With this in mind, we propose a research design that starts with exploration, making sure that we provide the necessary precautions regarding the four points Richards and King bring forward: privacy, confidentiality, transparency and identity protection. Ess (2006) advocates a practical view of doing research ethics. Researchers are perfectly capable of making ethical decisions within their own fields, assessing a variety of ethical considerations depending on the context. We therefore choose a perspective of ethical pluralism over dogmatism.

Reviewing the research process, we made an attempt to propose a way of implementing ethical reasoning as well as risk limitation and the safeguarding of personal data confidentiality. It must be emphasized that we want to ensure a maximum degree of academic freedom while identifying possible risks and limiting them.

Stage 1: Design

In this stage an idea will be turned into a research design. This process might start with a 'found data set' or a platform that triggers the researchers' interest and provides a starting point for possible research questions to be developed. The three elements listed in this stage in Figure 13.4 are therefore exchangeable and do not have to follow one upon the other. A topic will be combined with a possible forum or database. An inventory of the stakeholders regarding the information will be made. It lists the amount of personal information, the degree of vulnerability and possible risks such as confidentiality breaches. As a result, a decision will be made about which data will be scraped and how this will be done. This will raise issues concerning the terms of use of the data source, the quality of the

gathered information, the legal status and the feasibility of data collection. Researchers must argue why they are collecting data in a specific way. The process of data collection will be developed and data will be scraped accordingly. The risks of the next phase will be defined and limited as much as possible.

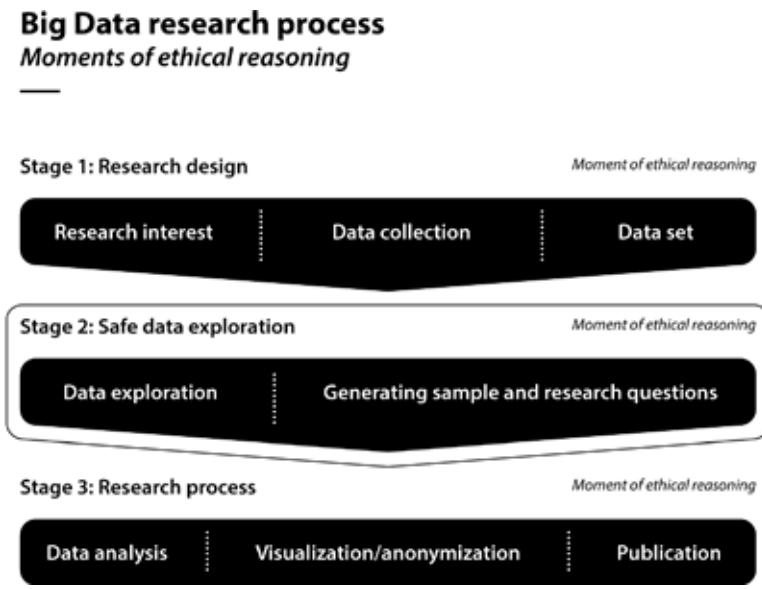
Stage 2: Safe Data Exploration

The second stage is an exploratory inquiry into the data set. It leads to the identification of patterns and samples and the formulation of a hypothesis. By exploring the data, researchers find out what the data is about and how it can be used. Several conceptual research questions and possible hypotheses are proposed. To conclude this stage, a definitive hypothesis is chosen with its corresponding sample. In this phase the data will be protected physically by using a stand-alone 'air gapped' computer. Prabhu (2015: 165) emphasizes that data usage needs to be governed tightly. Access to the data will be documented carefully, and only the necessary people will be allowed to work with the data. The data will be explored and filtered. Special attention will be given to patterns that might occur in the data. Research questions will be formed and a sample will be selected. At the end of this stage a decision will be made about whether a part of the data will be carefully anonymized and used in the third stage. Anonymization has to be processed carefully and must take into consideration the possibility of the existence of another data set consisting of partly similar data. The combination of two data sets has proved to be an effective method of de-anonymization (Narayana & Shmatikov 2006; Sweeney 2002). Completely wiping the data is also a possibility.

Stage 3: Research Process

The third stage involves testing the hypotheses which are formulated during the first stage. The use of the data now shifts from an exploratory environment to a research environment. The research should comply with the rules and ethical guidelines that are part of its specific scientific tradition and institution. If informed consent is stipulated by required guidelines, it should be requested. An opt-in or opt-out can be provided to people so they can actively make a choice about their data (Gleibs 2014; Prahbu 2015). Before possible publication, special attention will be given to the anonymization of sensitive data.

Fig. 13.4: Research process with safe data exploration.



Conclusion

Emerging new branches of humanities research dealing with the use of digital methods are raising questions about methods and ethics. Informed consent as the principle value affiliated with research ethics in the social sciences is under pressure due to two technological advancements: the rise of the internet and big data technologies. Informed consent has been an integral part of all guidelines concerning human subject research in the medical field and the social sciences. First, we demonstrated that the basis of these ethical guidelines in the Nuremberg Code, the Declaration of Helsinki and the Universal Declaration on Bioethics and Human Rights are from eras and discourses that have very little to do with big data research as it is currently being done. Although these guidelines can be very useful or even necessary in the final stage of a big data research project studying an SNS, in the second stage they would only limit the researcher.

Second, we showed that online user accounts and profiles are not equal to human subjects. Online profiles can better be seen as representations of people, not the people themselves, and, depending on the SNS being investigated, many users may provide fake or false information. Receiving

informed consent from the whole population of a social network or service is therefore unrealistic. And those who positively respond to a request might constitute a biased and unrepresentative sample. Another practical problem are the numerous inactive profiles online: a request for informed consent will not be answered by those who are no longer members of the community. We showed that in our own research this group would have amounted to close to 80 percent of the profiles. Again, expecting informed consent as a requirement for research to be ethical is unrealistic. This does not mean that researchers must not take all possible precautions to safeguard the confidentiality of the data collected.

To deal with the problems we described above we propose using a system of three stages in big data research on SNSs. Rather than favouring one ethical framework over another, we adopt a view of ethical pluralism, leaving it to the researcher to choose which to use, making appropriate reflections within their context. In the first stage a research design will be made, taking into consideration the stakeholders, type of data and a general direction of inquiry. After the gathering of data, in the second, exploratory stage hypotheses and samples are generated. Informed consent is not necessary in this stage, but since the nature of the data can still be very delicate, protection of the data is of the utmost importance. In the third stage, researchers have to adhere to the rules and guidelines that are mandatory in their specific field of research. In most social sciences informed consent is part of these guidelines and will therefore have to be respected. With this proposal we expect to catalyse both the philosophical and practical discussions about informed consent. To ensure that future research with new tools can be carried out in an ethical way, we need to experiment not only with methods but also with ethical frameworks. In order for us to find practices to protect research integrity we need to get our hands dirty.

References

Adkins, Lisa & Celia Lury. 2009. "Introduction: What is the Empirical?" *European Journal of Social Theory* 12 (1): 5.

Annas, George J. 1992. "The Changing Landscape of Human Experimentation: Nuremberg, Helsinki, and Beyond." *Health Matrix* 2 (2): 119.

Berlinguer, Giovanni & Leonardo De Castro. 2003. *Report of the IBC on the Possibility of Elaborating a Universal Instrument on Bioethics*. Paris: UNESCO.

Berry, David M. (ed.) 2012. *Understanding Digital Humanities*. New York: Palgrave Macmillan.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D.I. Kramer, Cameron Marlow, Jaime E. Settle & James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489(7415): 295-298.

boyd, danah & Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, And Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662-679.

Buchanan, E.A. & Charles Ess. 2008. "Internet Research Ethics: The Field and Its Critical Issues." In Himma, K.E. & H.T. Tavani (eds.), *The Handbook of Information and Computer Ethics*. John Wiley & Sons, 273.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner & Jefffrey Schnapp. 2012. *Digital_Humanities*. Cambridge, MA: The MIT Press.

Carberry, John. 2014. "Media Statement on Cornell University's Role in Facebook Emotional Contagion Research." *Cornell University Media Relations*. 30 June 2014. Retrieved from http://mediarelations.cornell.edu/2014/06/30/media-statement-on-cornell-universitys-role-in-facebook-emotional-contagion-research/.

Carlson, Robert V., Kenneth M. Boyd & David J. Webb. 2004. "The Revision of the Declaration of Helsinki: Past, Present and Future." *British Journal of Clinical Pharmacology* 57 (6): 695-713.

Christie, Bryan. 2000. "Doctors Revise Declaration of Helsinki." *B MJ* (321): 931.

Dörk, Marian, Sheelagh. Carpendale & Carey Williamson. 2011. "The Information Flaneur: A Fresh Look at Information Seeking." Proceedings of the SIGCHI Conference on Human Factors. *Computing Systems*: 1215-1224. ACM.

Dumas, Guillaume, David G. Serfass, Nicolas A. Brown, & Ryne A. Sherman. 2014. "The Evolving Nature of Social Network Research: A Commentary to Gleibs." *Analyses of Social Issues and Public Policy* 14.1: 374-378.

Ess, Charles. 2006. "Ethical Pluralism and Global Information Ethics." *Ethics and Information Technology* 8 (4): 215-226.

—. 2007. "Internet Research Ethics." In Joinson, Adam, Katelyn McKenna, Tom Postmes & Ulf-Dietrich Reips (eds.), *Oxford Handbook of Internet Psychology*. Oxford University Press: Oxford.

Gleibs, Ilka H. "Turning Virtual Public Places into Laboratories: Thoughts on ConductingOnline Field Studies Using Social Network Sites." *Analyses of Social Issues and Public Policy* 14.1 (2014): 352-370.

Grodin, Michael A. 1994. "Historical Origins of the Nuremberg Code." In *Medicine, Ethics and the Third Reich: Historical and Contemporary Issues*, edited by John J. Michalcyzk, 169-194. Kansas City, MO: Sheed and Ward.

The Guardian. 2010. "WikiLeaks embassy cables: the key points at a glance." Retrieved from www.theguardian.com/world/2010/nov/29/wikileaks-embassy-cables-key-points (accessed 2 January 2016).

Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer & Nicholas Christakis. 2008. "Tastes, Ties, and Time: A New Social Network Dataset using Facebook.com." *Social Networks* 30 (4): 330-342.

Manovich, Lev. 2011. "Trending: The Promises and the Challenges of Big Social Data." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 460-475. Mineapolis, MN: University of Minnesota Press.

Markham, Annette & Elizabeth Buchanan. 2012. "Ethical Decision-making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)". USA: Association of Internet Research. Sciences, 2nd edition, Vol 12. Oxford: Elsevier, 606-613.

Markham, A., & E. Buchanan. "Ethical Considerations in Digital Research Contexts." *Encyclopedia for Social & Behavioral Sciences* (2015): 606-613.

Michael, G.J. 2015. "Who's Afraid of WikiLeaks? Missed Opportunities in Political Science Research." *Review of Policy Research* 32(2): 175-199.

Narayana, Avind & Vitaly Shmatikov. 2006. "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)." Retrieved from http://arxiv.org/pdf/cs/0610105.pdf (accessed 24 February 2014).

Newitz, Annalee. 2015. "Is Cheater Site Ashley Madison Actually Growing by over a Million Users Per Month?" *Ars Technica*. Retrieved from http://arstechnica.com/tech-policy/2015/12/is-cheater-site-ashley-madison-actually-growing-by-over-a-million-users-per-month/ (accessed 2 January 2016).

Nuremberg Code. 1996 [1947]. "Permissible Medical Experiments." BMJ 1996 (313): 1448. Prahbu, Robinha. 2015. "Big Data? Big Trouble!" In *I nternet Research Ethics*, (ed.) Halvard Fossheim & Helene Ingierd, 157-172. Oslo: Cappelen Damm Akademisk.

Puschmann, Cornelius. & Engin Bozdag. 2014. "Staking Out the Unclear Ethical Terrain of Online Social Experiments." *Internet Policy Review* 3(4).

Richards, Neil M. & Jonathan H. King. 2014. "Big Data Ethics." *Wake Forest Law Review* 49: 393-432.

Rieder, Bernhard & Theo Röhle. 2012. "Digital Methods: Five Challenges." In *Understanding Digital Humanities*, ed. David M. Berry. London: Palgrave Macmillan. 67-84.

Rogers, Richard. 2009. *The End of the Virtual: Digital Methods*. Amsterdam: Amsterdam University Press.

—. 2013. *Digital Methods*. Cambridge, MA: The MIT press.

Schroeder, Ralph. 2014. "Big Data and the Brave New World of Social Media Research." *Big Data & Society* 1 (2).

Sweeney, Latanya. 2002. "k-Anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (05): 557-570.

UNESCO. 2006. "Universal Declaration on Bioethics and Human Rights." Paris.

White, Michele. 2002. "Representations Or People?" *Ethics and Information Technology* 4 (3): 249-266.

World Medical Association. 2013. "Declaration of Helsinki."

Zimmer, Michael. 2010. "But the Data is Already Public': On the Ethics of Research in Facebook." *Ethics and Information Technology* 12 (4): 313-325.

Zwitter, Andrej. 2014. "Big Data Ethics." *B ig Data & Society* 1 (2): 2053951714559253.