NEW FORMS OF DATA – LEGAL AND ETHICAL ISSUES

Marianne Høgetveit Myhren[1], **Vigdis Namtvedt Kvalheim**[1], **Audun Gabriel Løvlie**[1], **Katrine Utaker Segadal**[1,]

[1]**NSD – Norwegian centre for Research Data (Marianne.Myhren@nsd.no, Vigdis.Kvalheim@nsd.no, Audun.Lovlie@nsd.no, Katrine.Segadal@nsd.no)**

Introduction

Data is the currency of today's digital economy. Collected, analysed and moved across the globe, personal data has acquired enormous economic significance. According to some estimates, the value of European citizens' personal data has the potential to grow to nearly €1 trillion annually by 2020.[1] This paper will discuss legal and ethical challenges facing cross national research that rely on access to large scale data on an individual level. Data generated by governments and national statistical agencies, the use of internet and web-based services and the emergence of a digital economy, are also potentially valuable resources for researchers in the social sciences. However, their use is rather limited compared to their perceived research potential. The fact that these data originally and primarily are produced for purposes other than research, creates major concerns connected to the re-use of such data by scientists. In addition, the growing amounts of quantitative data and new ways of working with data in empirical social science research raise legal and ethical challenges.

The demand for stronger protection and harmonization of regulations and practices is high, particularly related to the use of data generated by or in relation to global communication networks. This has been an important driving force behind the new data protection regulation in the EU. The question is how the new legislation balances the interest in privacy against the use for historical, statistical and scientific purposes.

[1] http://europa.eu/rapid/press-release_MEMO-15-6385_en.htm

SERISS will focus on these issues particularly when linking new forms of data, e.g. administrative data, big data, social media data and biomarkers, with survey data. The impact of the new European General Data Protection Regulation (GDPR) and national data sharing practices, including that of national statistical agencies, will be described and examined. The aim is to develop strategic policy guidelines and best practices to address legal and ethical challenges.

New forms of data

The availability of new forms of data presents new challenges for research. However data such as administrative data and biomarkers are not strictly new forms of data, rather it appears that the mode or the extent of the collection or linkage of these data are rather new within the social science domain in most European countries (DASISH 6.1:7). The Nordic countries are the exception; they have a long tradition for research based on data from registries.[2] All the Nordic countries provide access for scientific research purposes. However data sharing across borders is still difficult even though the legal framework for data sharing is largely in place.

Big data and social media data on the other hand can be considered as new forms of data in scientific research. The term "Big data" can be defined as a term for data that is of such a scope that more processing power than normal is required in order to collect and analyse it. Social media data can be distinguished from "Big data," as the content is user generated and intentionally and actively shared. The content can be shared privately (limited to favourite friends lists, closed circles) or publicly. Currently, researchers are making use of social media data to derive quality insights, for presenting purposes and for recruitment purposes in survey research (Murphy et.al 2014). Smith and Kim argues that combining detailed survey data with geographically- and sociologically based data gives context to people's lives and give researchers a deeper understanding of the ways in which individuals and societies function (Smith and Kim 2013). Information on both personal level and aggregate level are of interest. This increased interest in using and/or linking administrative data, big data and data from

[22] The Nordic countries are particularly well known for their large number of registries covering the entire population or significant subpopulations. The registries are established for administrative and statistical purposes and can be merged by personal identification numbers to form event history data bases that can be updated every year. The merged registries are used for production of statistics and as well as research purposes.

social media platforms like Facebook and Twitter to survey data, raises serious challenges when it comes to the protection of privacy, informed consent and data confidentiality. This in turn increases the necessity to focus on legal and ethical challenges.

Current and new legal framework

The current legal framework is based on Directive 95/46/EC (Directive) that was introduced in 1995. Since then there have been significant advances in information technology and fundamental changes to the ways in which individuals and organisations communicate and share information. In addition the various EU member states have taken divergent approaches to implementing the Directive, creating difficulties for the free flow of personal data across Europe. In January 2012 the European Commission put forward its EU Data Protection Reform with the aim to make Europe fit for the digital age, put an end to the patchwork of data protection rules that currently exists in the EU, unlock opportunities and remove unjustified barriers which limit cross border data flow. It also aims to ensure a consistent and high level of data protection to provide legal certainty and trust. In April 2016 the General Data Protection Regulation (GDPR) was formally adopted by the EU Parliament and Council and will be applied after a 2 year transition period. The GDPR is a regulation, not a directive. This means it applies directly to all EU and EEA countries, without the need for national legislation. According to the EU, the Commission will work together with the member states and the Data Protection authorities - the future European Data Protection Board (EDPB) – to ensure a uniform application of the new rules. An essential part of establishing a uniform application is a uniform implementation of the term personal data.

The scope of the GDPR - What is Personal Data?

The Directive defines personal data as “any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity; (Directive, Article 2 (a)). The definition of personal data in the GDPR is similar to the definition in the Directive, but location data, online identifiers and genetic factors have been included in the definition of personal data in the new law (Article 4 (1)). In addition, genetic data and biometric data are recognized as sensitive data requiring extra protection.

To determine whether a person is identifiable or not Recital 26 states that one should take account of all the means reasonably likely to be used, either by the controller or by another person, to identify the person directly or indirectly. Other factors, such as the cost of and the amount of time required for identification, the available technology at the time of the processing, and technological development also need to be taken into account and considered.

The current definition of personal data is implemented across Europe with various degree of strictness. UK and the Netherlands for instance have a liberal interpretation. While countries like Norway and Estonia have a strict one. As a result of this more research projects fall outside the scope of the privacy regulation in the UK and Netherlands than in Norway and Estonia (DASISH 6.5). The definition of personal data as stated in Article 4 will probably lower the threshold for considering information as directly or indirectly identifiable, resulting in more research projects falling within the scope of the law, at least in countries which have a liberal definition today. In countries with the stricter definition, this could mean that more projects could fall outside the scope of the privacy regulations and be considered anonymous. A standardised practice and implementation of the definition of personal data can therefore improve the possibilities for data access, data sharing and in turn improve conditions for cross national research.

The GDPR – Continuity more than change

The general impression of the GDPR is that it will not lead to dramatic changes for European research institutions or statistical agencies. The new law aims at eliminating fragmentation and providing consistency and coherence for the whole of the Union. There has for instance been a lot of effort put around the creation of a so-called One-Stop-Shop that will streamline cooperation between the data protection authorities on issues with implications for all of Europe. Greater harmonisation and standardisation in the field of data protection should benefit the research sector as it will make it easier to exchange data and promote and stimulate cross national research collaboration.

Furthermore the GDPR has turned out to be fairly research-friendly as the most important special provisions from the Directive are continued, clarified and strengthened. Amongst

other this applies to the possibility to process personal data based on other grounds than consent.

Generally the processing of personal data for purposes other than the original one is only legal when the data subject consents. However the GDPR has specific provisions on processing of personal data for health purposes and for historical, statistical and scientific research purposes. When listing principles relating to processing of personal data Article 5 (b) and Recital 50, states that "further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89 (1), not be considered to be incompatible with the initial purposes ('purpose limitation'). Also the processing of sensitive personal data is allowed if the data subject has given explicit consent **or** the processing is necessary for research purposes (Article 9 (a) and j)).

These provisions are essential and allow for secondary use of data collected from administrative registers or the internet, for scientific research purposes without having to obtain further consent from the individuals, if appropriate safeguards are in place. All legislation intends to serve a normative function. Thus this provision clearly signalizes to citizens that research is in a unique position, a specific legitimate activity and that society recognises the role of research as a producer and communicator of knowledge. Unambiguous wording with regards to the status of scientific, statistical and historical purposes within the privacy regulation is the best way to establish that privacy protection and access to personal information for scientific purposes can go hand in hand. This should also give a clear signal to the institutions that over time will give substance to the law and stipulate the conditions for storage, use, alignment and disclosure of personal data, such as statistical offices. However, the GDPR also provides some scope for Member States to implement local rules on certain matters, for instance on the use of health data (Article 9, 4). Additionally it entails certain conditions for the processing to be legal.

Conditions for processing of personal data

The processing of personal data for scientific research and statistical purposes has been given a unique position in the regulation's general provisions. However these provisions also entail specific conditions to ensure the rights and freedom of the data subject. The GDPR promotes techniques such as **anonymisation** (removing personally identifiable information where it is not needed), **pseudonymisation** (replacing personally identifiable material with artificial identifiers), and **encryption** (encoding messages so only those authorised can read it) to protect personal data. This practice is already standard in the Nordic countries. However, in some countries like the UK, these requirements imply stricter conditions and cause some concerns related to the possibility of conducting cohort studies.

Similarly to the Directive, individuals must be provided with certain information that explains the context and purpose for the use of their personal data. However, the GDPR expands the list of what individuals need to be told, such as whether data will be transferred and how long it will be stored. Similar information must be provided to individuals if the data is not collected directly from the data subject. However, the obligation to inform individuals does not apply if providing the notice is likely to render impossible or seriously impair achieving the objectives of the research project. When assessing whether the exemption provision applies, consideration should be given to the number of data subjects, the age of the data, and what measures are implemented to protect the legitimate interests of the data subject (Article 14, 5(b) and Recital 50). This is in line with current practice but the GDPR introduces an obligation for the controller to provide all information to and communication with the data subject in clear and plain language adapted to the data subject, particularly in relation to children (Article 12, Recital 58).

The GDPR also includes a reinforced "right to be forgotten". However it also protects freedom of expression and the freedom of the media, as well as historical and scientific research. It provides exemptions for these sectors asking Member States to adopt national laws to guarantee the respect of these fundamental rights. This allows archives to continue

operating on the basis of the same principles as today. In short, the right to be forgotten is not absolute and does not affect scientific or statistical purposes.[3]

The GDPR strengthens the data protection officer arrangement and in practice makes it mandatory for most of the research sector as well as all national statistical agencies (Article 37). The general notification obligation and the obligation to obtain a licence from the national data protection authority are revoked in its entirety. Instead all controllers must carry out an impact assessment of the envisaged processing operation in relation to the protection of personal data together with a data protection officer (Article 35, 1 and 2).

The data protection officer will be a main element in the system for regulating, controlling and documenting the processing of personal data for different purposes. Hence the data controllers (the research institutions) will get more responsibility and the data protection officials' role will become mandatory and expanded. Data protection authorities will keep their supervisory role, and will be given more power. In addition a new European Data Protection Board (EDPB) will play a greater role with wider role with ensuring a uniform application.

New forms of data – a challenge for the data subjects' confidentiality?

In Opinion 7/2015 – Meeting the challenges of big data – the European Data Protection Supervisor (EDPS) argues that the application of big data raises serious concerns for individual's right to privacy. The OECD Global Science Forum claims that there is a need for greater transparency when using new forms of data in research, maximizing the gains in knowledge derived from such data while minimizing the risks to individuals' privacy, seeking to retain public confidence in scientific research which makes use of new forms of data.[4]

Techniques such as anonymisation of personal data have multiple purposes within research: as a marker of ethical practice, a means of reducing the need to get prior approval from ethical boards/data protection officers/supervisors, and as a safeguard for protecting

[3] http://europa.eu/rapid/press-release_MEMO-15-3802_en.htm

[4] OECD Global Science Forum (2013): «New Data for Understanding the Human Condition: International Perspectives», page 2.

respondent privacy. However, the growing capabilities of technology to gather and analyse data have raised concerns over the potential reidentification of anonymised datasets (Nunam and Di Domenico 2016). New computational techniques can identify people or trace their behaviour by combining just a few snippets of data. "Privacy as we have known it is ending, and we're only beginning to fathom the consequences" (Enserink and Chin 2015). According to a famous study of the 1990 census data in the US, 87% of the US population could be identified by their zip code, combined with gender and date of birth (Ohm 2010). In January 2013 researchers were able to identify individuals and families from supposedly anonymous DNA data from publicly accessible genealogy databases (Gymrek et al. 2013: 321-324).

Couzin-Frankel argues that scientists, as a result of an increased risk of reidentification, no longer can guarantee privacy for research objects and that they are looking for new ways to build trust. She shows how medical researchers have addressed this issue by giving research volunteers a louder voice in research and transparency about how the research is conducted (Cuzin-Frankel 2015). De Montjoye et.al suggest that large scale data sets of human behaviour can transform how we perform research, and argue that US and EU privacy laws are inadequate for meta data sets. Therefore they suggest that from a technical perspective their results emphasize the need to move, when possible, to more advanced, and probably interactive individual or group privacy conscientious, technologies (De Montjoye et.al.2015*)*.

One of the aims of the GDPR is to build trust by assuring a high level of privacy protection, while also enforcing stricter obligations with regard to data security. Article 25 promotes data protection by design and default. By building data protection into the design, and adjusting data protection to allow transparency and more user control, the EDPS claim that controllers may benefit from big data while at the same time ensuring individuals dignity and freedom. The EDPS believes that sustainable and responsible controllers of big data must rely on four essential principles: transparency, high degree of control, design user friendly data protection products and accountability (Opinion 7/2015).

New forms of data – Ethical challenges

Technological developments and in particular the use of big data and social media data for

research purposes pose several ethical challenges when conducting research online.[5] An illustrating example is the fact that information made publicly available is exempted from the obligation to obtain consent both in the Directive (Article 8 (e) and the new GDPR (Article 9, 2 (e)). Hence a research project that includes data made publicly available on Facebook can legally use these data without consent. Another ethical dilemma is the issue of data determinism, a development where more and more analyses of e.g. big data may result in us not being judged on the basis of our actual actions, but on the basis of what all the data about us indicate our probable actions may be (Datatilsynet 2013:7).

According to Descombe, ethics concern what 'ought' to be done. Research ethics seek to protect the interest of the research participants and are guided by the core principles of protecting the rights of free will, privacy, confidentiality and well-being of research participants, and minimize the burden of participation to the greatest extent possible (Denscombe 2002:175-176). In the guidelines for ethical decision making and internet research the Association of internet researchers (AoIR) states that individual and cultural definitions and expectations of privacy are ambiguous, contested, and changing. "People may operate in public spaces but maintain strong perceptions or expectations of privacy. Or, they may acknowledge that the substance of their communication is public, but that the specific context in which it appears implies restrictions on how that information is -- or ought to be -- used by other parties" (AoIR 2012).

Hence even though a project may fall outside the scope of the privacy regulation it's important to remember that ethical guidelines still applies. Marika Lüders argues that it's important to have in mind that the concept of privacy online is a moving target, constantly being negotiated and renegotiated as a consequence of how we perceive the boundaries between public and private spheres. As a consequence of this she proposes that we need to have a processual approach to research ethics, because particular judgement made in different case or different stages of the studies cannot be easily applied in other research cases (Lüders 2015:77-95).

[5] In Norway, the ethical guidelines for research stress the importance of the researcher to consider people's perceptions of what is private and public. Reader debates in online newspapers are for instance considered public, while personal blogs may be available for everyone to read, but can be regarded as personal space by the author (NESH 2014).

Concluding remarks

The general impression of the new regulation is that it is research friendly and that it safeguards the interests and the needs of scientific research institutions and statistical agencies. Most of the important special provisions for the processing of personal data for research and statistical purposes have been continued, clarified and strengthened. Most importantly is the fact that the processing of personal data for scientific research or statistical purposes, is always compatible with the initial purposes. A standardised practice and implementation of the definition of personal data can also improve the possibilities for data access, data sharing and in turn improve conditions for cross national research. Hence the legal basis for using new types of data in a social survey context, including big data, biomarkers, social media data and administrative data, is in place, *but* the possibility of member states to maintain or introduce further conditions, including limitations, with regard to the processing of genetic data, biometric data or data concerning health may still pose a challenge for harmonization across Europe.

Even though the legal basis for using new forms of data is largely in place, the use of these data in conventional or new ways raise ethical challenges that requires a dynamic approach. This is especially relevant regarding how to perceive the boundaries between public and private spheres online. The increased risk of re-identification when using these new forms of data creates a need for greater transparency in order to retain public confidence and trust in the unique position of scientific communities as producers and communicators of knowledge.

References:

Association of Internet Researchers (AOIR) (2012) "Ethical Decision-Making and Internet Research", Recommendations from the AoIR Ethics Working Committee (Version 2.0).

Couzin-Frankel – "Trust me I'm a medical researcher", *Science* 30 Jan 2015, Volume 347, Issue 6221, pages 501-503. Available at: http://science.sciencemag.org/content/347/6221/501.full?intcmp=collection-privacy

DASISH Deliverable 6.1 (2013): "Report about New IPR Challenges" DASISH Work Package 6:"Legal and Ethical Issues", Contributing Partners and Editors: Schmidutz, D., Ryan, L. Gjesdal, A., De Smedt, K.

DASISH Deliverable 6.5 (2014) : "Handbook on legal and ethical issues for SSH data in Europe, Part II", Contributing Partners and Editors: Bøe, M., Rød, L. Parra, C., De Smedt, K., Kvalheim, V., Segadal, K., Kvamme, T., Dione, B., Samdal, G.

Datatilsynet (2013). "Big Data – privacy principles under pressure". Available at: https://www.datatilsynet.no/globalassets/global/04_planer_rapporter/big-data-engelsk-web.pdf

De Montjoye, Y-A. et.al. (2015):"Identity and privacy". *Science*, Volume 347, Issue 6221, pages 536-539. Available at: http://science.sciencemag.org/content/347/6221/536.full?intcmp=collection-privacy

Denscombe, M (2002): *Ground Rules for Good Research: A 10 point guide for social researchers*. Open University Press: Buckinham.

DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL (1995): http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:PDF

EDPS - European Data Protection Supervisor (2015) – "Opinion 4/2015 - Towards a new digital ethics – Data, dignity and technology". 11 September 2015. Available at: https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-09-11_Data_Ethics_EN.pdf

EDPS - European Data Protection Supervisor (2015) – "Opinion 7/2015 - Meeting the challenges of big data – A call for transparency. User control, data protection and accountability." 19 November 2015. Available at: https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf

Enserink, M and Chin, G. (2015) "The End of privacy". *Science* 30, Volume 347, Issue 6221, pages 490-491. Available at: http://science.sciencemag.org/content/347/6221/490.full?intcmp=collection-privacy

European Commission (2015) – "Fact Sheet - Questions and Answers - Data protection reform" (online). Available at: http://europa.eu/rapid/press-release_MEMO-15-6385_en.htm

General Data Protection Regulation (2016): [online] Available at: http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf

Gymrek,M. et.al. (2013). "Identifying Personal Genomes by Surname Inference", *Science* Volume 339, 321–324.

Ohm, Paul - 'Broken promises of privacy: responding to the surprising failure of anonymisation', UCLA Law Review 2010 and 'Record linkage and privacy: issues in creating new federal research and statistical info', April 2011

Murphy, Joe et.al (2014). "Social Media in Public Opinion Research" - Executive Summary of the Aapor Task Force on Emerging Technologies in Public Opinion Research [online] Available at: https://poq.oxfordjournals.org/content/early/2014/11/24/poq.nfu053.full

National Committee for Research Ethics in the Social Sciences and the Humanities (NESH)(2014). "Ethical Guidelines for Internet Research". Available online: https://www.etikkom.no/globalassets/documents/english-publications/ethical-guidelines-for-internet-research.pdf

Nunan, Daniel and Maria Laura Di Domenico (2016): "Exploring reidentification risk; Is anonymisation a promise we can keep?" *International Journal of Market Research* Vol. 58 issue 1:19-34

OECD Global Science Forum (2013): «New Data for Understanding the Human Condition: International Perspectives».

Smith, Tom W. and Jibum Kim. 2013. "An assessment of the Multi-level Integrated database Approach." ANNALS of the American Academy of Political and Social Science 645:185–221.