OECD publishing

Please cite this paper as:

OECD (2016-11-08), "Research Ethics and New Forms of Data for Social and Economic Research", *OECD Science, Technology and Industry Policy Papers*, No. 34, OECD Publishing, Paris.
http://dx.doi.org/10.1787/5jln7vnpxs32-en

OECD Science, Technology and Industry Policy Papers No. 34

Research Ethics and New Forms of Data for Social and Economic Research

OECD

This paper was approved and declassified by the Committee on Scientific and Technological Policy on 12/08/2016 and prepared for publication by the OECD Secretariat.

Note to Delegations:
This document is also available on OLIS with the cote:
DSTI/STP/MS(2016)2/FINAL

ACKNOWLEDGMENTS

This report is the work of an Expert Group, appointed by the OECD Global Science Forum (GSF), bringing together a wide range of expertise from many countries. The Group has worked over a two-year period to address what is rapidly becoming one of the most pressing challenges for social scientists – research use of new forms of data. This is not just a question of the legality of such usage, but of 'doing the right thing'. While some countries and agencies are beginning to tackle this problem, the global nature of many new forms of data requires a more concerted international effort. The report presented here is a step in this direction. It is advisory, but it carries the weight of full support from all members of the Expert Group and has been endorsed by GSF and the OECD Committee on Science and Technology Policy.

In carrying out our work, the Expert Group has had financial support from the science ministries in the countries that nominated members to the group, from the OECD Global Science Forum secretariat, and from the two agencies that provided dedicated project and conference funding – the UK Economic and Social Research Council and the US National Science Foundation. This support is gratefully acknowledged.

The full membership of the Expert Group is shown at Appendix 1. We would like to thank all members of the Group for the time and expertise they contributed to the production of this report. To facilitate drafting of the final report a smaller editorial group was formed, comprised of members from South Africa, Norway, the United States of America, and the United Kingdom. This group has worked carefully to ensure that the document reflects the views of the Expert Group and achieves the high standards required of OECD publications. Special thanks are due to Carthage Smith at the OECD Global Science Forum and Dr Deborah Colson, formerly Chief Scientific Officer for Life Study at University College London, who provided clarity and impetus when it was most needed.

Peter Elias, Chair of the Expert Group

Hallvard Fossheim, Vice Chair of the Expert Group

PREFACE

New kinds of data are rapidly becoming available in massive quantities, providing a record of the transactions we carry out, the communications we make and other social and economic activities. Although such data are not collected primarily for research, they offer considerable research potential and have the capacity to yield improved insights into society, health, the economy and political behaviour. However, along with the potential benefits, the availability of these data in a rapidly changing digital environment also presents a number of ethical challenges. These include risks relating to the disclosure of the identities of individuals or organisations; reputational risks for organisations collecting or creating data; and issues around the ethics of research using these data.

In 2013, the OECD published a report on 'New Data for Understanding the Human Condition'[1], which recommended the development of an internationally recognised framework code of conduct covering the use for research of new forms of personal data. Subsequently, the OECD established an Expert Group (see **Appendix 1**) to consider the ethical issues that may arise from research use of new forms of data, with particular reference to the balance between the social value of research using such data and the protection of the well-being and rights, including privacy rights, of individuals; and to draft guidance on an ethical approach to the use of such data for research[2]. This is the report of that Expert Group. This report also builds upon a number of related initiatives by the OECD and other international and national bodies, including *inter alia*: guidelines relating to the protection of privacy and trans-border flows of personal data; access to research data from public funding; international collaboration on data access; health data governance; and ethical protocols and standards for research in the social sciences.

While we are aware of other work that has sought to address ethical issues relating to the research use of New Forms of Data, this report breaks new ground in a number of ways. First, it re-examines the principles underpinning research ethics in the context of the changing research environment. Second, the issue of 'what is legal' in different countries is considered as well as 'what is ethical', seeking to determine the adequacy of national legislation to safeguard the interests of data subjects. Third, while some countries are well-placed to provide the framework for the relevant research ethics, others are just beginning to take steps in this direction. The Expert Group included members from a wide range of countries, ensuring that there was an emphasis on the sharing of knowledge and expertise in this area. Finally, the recommendations are structured to be equally applicable to private and public sector research.

This report provides a set of high-level recommendations that can be used to underpin a system for the ethical governance of research, applicable at all stages of the research process from proposal formulation to publication. Many OECD countries already have systems in place to provide research governance/oversight, although they are normally used for public sector research. Where these already exist, the majority of the recommendations in this report (if not already in place) could be enacted with few significant cost implications. Where no such systems exist or are incomplete, research funders or relevant national oversight bodies will need to weigh the potential costs associated with ensuring adequate governance for research using New Forms of Data against the possible harm that could arise from the lack of a research governance framework.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Overview

The variety and volume of the New Forms of Data that are available with potential to inform research in the social sciences is rapidly expanding. Much of this falls in the category of 'Big Data' which are characterised by their size and complexity and the fact that they are often not amenable to the more traditional forms of statistical analysis used in social science research. However, size is not everything, and New Forms of Data can include relatively small datasets. New Forms of Data that are of interest for social sciences research include a diverse range of data and data sources, such as information on internet usage and commercial transactions, data derived from tracking systems, and government information (see **Appendix 2**). In this new era for research, traditional distinctions between public and private data, or public and private use, are increasingly blurred as massive amounts of data on the web are generated by the public but theoretically owned by companies. These data increasingly cross national (or disciplinary) boundaries which can add to the complexity of the issues.

Research using New Forms of Data, which is often collected for purposes other than research, raises novel and important ethical issues. These issues are evolving as technology progresses and cannot be fully addressed by legislative action. There is a need for robust ethical guidance that enables important research to continue and at the same time safeguards the interests of all parties. Because of the novelty and complexity of the issues raised by New Forms of Data, there is a need to re-visit fundamental philosophical and ethical principles and develop guidance for all those involved in the social science research endeavour. In order to be effective, such guidance must complement existing legal and regulatory frameworks.

This report summarises the discussion and recommendations of an international Expert Group relating to: ethical principles; governance; legal frameworks; and critical topics such as privacy, consent, anonymity, commodification of data, data sharing and security, and public engagement. In addition to guidance and recommendations targeted at specific stakeholders, the report identifies a number of opportunities for facilitating good practice in the era of New Forms of Data that require collective agreement and action (See **Section 2, Table 2**).

1. Ethical principles

The general ethical principles relating to social research fall into two broad groups. The first concerns how researchers relate to each other and their own methodological standards. Of key importance is *constructively critical openness*. The second group concerns how researchers relate to the surrounding world, primarily the data subjects. Here, the key principles are *respect*, *ensuring good consequences*, and *justice*. In considering these principles in relation to research using New Forms of Data, the Expert Group identified a set of eight basic rules for the ethical collection, sharing, and research use of New Forms of Data, and would encourage their consideration by all stakeholders involved in the research process:

1. Mechanisms for the safe and responsible sharing of personal data, including mechanisms for the protection of privacy of data subjects as well as for public input and accountability, should be

established and made public by data owners/controllers. Data should be shared as openly as is feasible within the relevant legal and ethical constraints.

2. The default position should be that personal data is not collected, processed or shared without informed consent. Efforts to update consent for new and unanticipated uses should be made where feasible.

3. Clear articulation of purpose should be provided before a research project using personal data is carried out. In many instances, this will entail the development of transparent long term plans and mechanisms for communicating any updates.

4. With a view to both the impact of the research and respect for data subjects, data quality should be considered to ensure that it is fit to fulfil the stated research purpose.

5. Before a research or data collection project is undertaken, care should be taken to consider potential negative impacts, for individuals or groups, arising from the proposed project. Any potential negative consequences should be weighed against societal benefits, taking account of any mitigating actions to reduce the risk or impact of potential negative consequences.

6. Unambiguous distribution of responsibilities should be agreed in advance of any research-related data handling.

7. Data holders, research funders and researchers have a responsibility to consider how their role in a proposed research project would contribute to the balance of power and influence between their institutions and individual data subjects.

8. Data holders and research institutions should ensure they have access to an ethics review body (ERB) with the capability to review proposals to use New Forms of Data for research.

Respect for these basic rules will ensure that research using New Forms of Data is conducted in an ethical manner. However, they do not provide the practical guidance that is required by those who approve and fund research, those who collect and hold data with research potential and researchers. The report spells out specific recommendations that provide guidance to these groups. The report does not prescribe how these recommendations should be implemented. What they do provide is a framework for implementation.

2. The governance of research using New Forms of Data

The advent of New Forms of Data has created new opportunities for data collection, secondary analysis, and experimental interventions, for both established actors and those who have not been part of a traditional research governance system in the past. There is a need to adapt existing systems and or establish new governance arrangements in order to encompass both traditional and new research actors. This is already happening in some countries and settings and there is considerable potential for exchanging and implementing best practices.

Institutional Review Boards or Research Ethics Committees (here referred to collectively as Ethics Review Bodies or ERBs) are a key element of established research governance systems. Although ERBs may have powers to withhold or withdraw approval from research projects, they operate in an ‘honour’ system where the nature of institutional support they receive may influence whether they are able to provide optimal independent oversight. ERBs will continue to have an important role to play in the era of New Forms of Data, provided that they are adequately supported and have the necessary independence and expertise to carry out their function.

Recommendation 1*: all stakeholders with an interest in the use of New Forms of Data for research should implement measures, based on current best practices, to strengthen the governance of research using New Forms of Data (see Section 2, Tables 1 and 2).*

Recommendation 2*: research funding agencies, research organisations and researchers should ensure that any research proposing to use personal data is referred for review by an appropriate independent ethics review body.*

3. Legal frameworks

An ethical approach to the use of New Forms of Data for research purposes helps determine whether and how research should be undertaken, whereas the legal frameworks surrounding research data specify what must or must not be done to comply with relevant laws.

Some countries have legal instruments to ensure that the potential public benefits arising from research using personal data are weighed against any compromise of privacy rights. This 'public interest' test is particularly important where consent from data subjects has not been obtained. In such cases, or where the research is based on New Forms of Data that cross national boundaries (and therefore jurisdictions), ethics review bodies will normally be well placed to consider the specific risks and benefits.

Recommendation 3: ***National and multi-national research funding agencies*** *should ensure that researchers have shown in their research plan that they:*

- *are cognisant of the relevant legal frameworks that may impact upon their access to and use of personal data for research;*
- *understand the adequacy of such legislation to protect the privacy of data subjects; and*
- *understand their legal responsibilities in relation to data collection, storage, processing, and sharing*

Recommendation 4: ***Ethics review bodies*** *(ERBs) should ensure that their policy and practice can encompass the assessment of respect and privacy issues in proposals for data access and sharing where existing legal frameworks may not provide adequate protection for the data subjects, or where the data and/or research cross national boundaries. In the latter case, the ERB may need dialogue with the relevant researchers and/or ERBs in those countries. Recommendations made by the ERB in this respect should be incorporated into the research design and reflected in any subsequent use of the data collected*

4. Addressing specific issues

4.1. Privacy

The three basic ethical principles of respect for persons, due consideration of good/bad consequences and justice are all relevant to research using personal data. Privacy plays into each of these three principles. It is important to keep in mind privacy's central role in enabling people to define, develop, and maintain their personal and social identities. Privacy is also central to the ability of an individual to participate fully in civil and economic life without retribution or discrimination. The notion of 'privacy' is complex and context-dependent. Due to this complexity, a simple recommendation on best practice in the use of New Forms of Data for research in relation to privacy cannot be formulated. A model that may be helpful in considering privacy issues is set out in **Appendix 4**.

Recommendation 5: ***Researchers and those involved in reviewing research proposals*** *should consider privacy protection, recognising that this is a complex issue with both legal and ethical aspects. For each proposal there should be a plan for clear communication to relevant audiences on how their privacy will be protected in research using personal data.*

4.2. Consent

At the core of what is deemed an ethical approach to research using personal data is the concept of 'informed consent', where the individual whose data is collected is informed about the purpose of the research and consents to the use of their data for these purposes. Much of the power of analysing New Forms of Data, however, is likely to lie in the later re-use of data in ways that the collector had not anticipated. Obtaining informed consent may in some cases not be possible, and protection against identification of a data subject may require more than operational methods alone can provide. Even when consent is possible there may be significant limitations relating to: the nature of the consent given and the time period for which it is valid; the capacity of the researchers to convey the balance between personal risk and potential societal benefits; and the balance of power between the data controller and the data subjects. Therefore, informed consent as a one-off event may not always be a feasible way to ensure respect for those involved as data subjects. It may be timely to consider new approaches to respecting the rights of those who provide personal data.

Better public engagement, informing those who may be affected about the possible risks and potential benefits from proposed research, can help to address the limitations of informed consent in the era of New Forms of Data. Above all else, as new approaches to communication and consent with research subjects are developed and tested, there is a need for a more harmonised approach to the regulation of procedures for consent; for the governance of research where informed consent is not an option; and for the maintenance of data protections designed to prevent disclosure of identities.

Recommendation 6: ***researchers*** *should:*

- *for any research plan (whether it includes collecting new data or uses previously gathered data) produce and make available a brief statement understandable to non-experts, explaining the general purposes and motivations for the research, together with an assessment of the potential risks to individuals or groups associated with the data to be used for research;*
- *consider the means of obtaining, and wording of, the consent sought for new data collection with a view to 'future-proofing' the consent to enable future research projects to use the data and, where possible, offer the public and in particular research participants the means to receive updates about the progress of the research, including previously unanticipated uses of data and opportunities to reaffirm consent for use, where applicable.*

Recommendation 7: ***data controllers, research funding agencies, ethics review bodies and researchers*** *should give careful consideration to the nature of any consent already obtained or required for the processing of personal data for research. Is it valid for the specified research? If not, can consent be obtained?*

Recommendation 8: ***ethics review bodies*** *should, where consent for research use of personal data is not deemed possible or would impact severely upon potential research findings, evaluate the potential risks and benefits of the proposed research. If the proposed project is deemed ethically and legally justified without obtaining consent, ethics review bodies should ensure that information is made publically available about the research and the reasons why consent is not deemed practicable, and should impose conditions that minimise the risk of disclosure of identities.*

4.3. Anonymity and de-identification

Social and economic researchers rarely need access to the identity of data subjects; their concern is with the relationship between variables within the data they seek to analyse. If the identity of an individual or an organisation cannot be inferred from a specific dataset made available for research, because identity markers have been replaced with an identifier which has no meaning to anyone but the data owner, the dataset is considered 'de-identified'. This means that, in principle, there should be virtually no risk of the disclosure of identities provided that the data are made available without any further linkage or data

matching being allowed. However, in era of new and big data, it is increasingly likely that different datasets could be matched and overlaid. De-identifying data simply by removing direct identifiers cannot therefore provide any confidence that the identity of a data subject is protected.

Many communities and/or countries now use a 'safe setting' or secure data enclave for data access, matching and/or analysis. The concept of a safe setting as a means of protecting the privacy rights of individuals requires mechanisms and security measures which must include the application of five strict (responsible and ethically robust) conditions. These conditions - often termed the 'five safes': safe people; safe projects; safe data; safe environment; and safe outputs - are designed to minimise the risk of disclosure. Although potentially applicable to New Forms of Data, experience with operationalising these five conditions for New Forms of Data is currently limited. Further work on preserving anonymity and protecting privacy is needed.

Recommendation 9: ***research funding agencies*** *should encourage further research on the development of statistical methods and software to provide assurances that the privacy of subjects in research using New Forms of Data is maintained, well understood and easy to implement.*

Recommendation 10: ***data holders, research funding agencies, and researchers*** *should share best practices in the creation and operation of safe settings, ensuring that restrictions on accessibility are minimised whilst maintaining data security.*

4.4. Ethical aspects of the commodification of data

The commodification of data refers to the buying and selling of personal data that have been given by individuals, often in another context, and perhaps in exchange for services. This generates challenges relating to the basic research ethics principles. These challenges can be compounded by a lack of clarity about the real ownership of the data.

The commercial sector is often the creator of datasets. Many of these are compiled for analysis aimed at gaining insights useful for business, for example insights into consumer profiles or behaviour. In some cases, companies may make this data (or access to algorithms that were developed from the data) commercially available. Since New Forms of Data and information on the internet are often available for wider use and analysis by others, they can easily be exploited directly or indirectly by the market economy through large internet service providers who have the ability to build large databases by borrowing from various sources. Whilst this is not a problem as such, issues arise when the commodification and use of information and research data by one agent precludes its use by others.

There is increasing collaboration between the public and commercial sectors due to the potential for mutual benefit in ensuring the availability of high-quality data for scientifically and ethically sound research. There is scope for these partnerships to facilitate good practice, on issues such as consent and privacy, and for the sharing of new methodological approaches. Shared solutions also need to be developed on ownership and licensing of data and other intellectual property generated through these partnerships. It is essential that high ethical and scientific standards are maintained by all parties in order to generate and retain public trust.

Recommendation 11: ***For data holders, research organisations, and researchers:*** *Where personal data are bought or sold on a for-profit basis to inform research, information about the nature of these transactions should be included with reported research results.*

4.5. Data sharing and security

Curation activities (which include the provision of access to data) should minimise the risks for identification of subjects from data, and should maximise the opportunity for research. Any framework for the responsible use of New Forms of Data must explicitly cover the responsible sharing of those data beyond the original research team. The responsibilities of all third parties involved in maintaining data (including trusted third parties which may link different datasets) must also be clear, and the risks posed by any combination or linkage of data need to be taken into consideration.

There are four key controls required for personal data to be successfully curated and amenable to re-use: consent, privacy rights, ownership (*e.g.* copyright, intellectual property), and research integrity. For curation to be successful when applied to New Forms of Data, each of the four controls (consent, privacy, ownership, research integrity) needs to be applied in different ways.

Recommendation 12: ***researchers and data holders*** *should:*

a. *establish guidelines and mechanisms through which applications to access data under their control may be made for publically-funded research;*
b. *ensure that the requirements and processes for researchers to access the data under their control are made publically available;*
c. *evaluate the potential for re-identification of individuals when depositing new putatively de-identified or putatively anonymised datasets, releasing datasets, adding new data, or developing data access platforms.*

4.6. Public engagement

New findings from social science research have potentially important and beneficial implications for the public, including research participants. However, there is a risk that any perceived invasion of individual privacy or cases of unreasonable surveillance could generate heated controversy and, with time, undermine the social science research enterprise. Public engagement is essential for researchers to reach better ethical solutions on difficult issues and to ensure public acceptance and trust of such research. Directly communicating information about projects and their rationale to the public may reduce the risk of public distrust or outrage. The 'public' for pragmatic purposes could be the attentive public and opinion leaders that would be at the centre of any controversy.

There are a number of approaches to formal democratic deliberation. One useful approach is 'Deliberative polling', summarised in **Section 4.6** together with a discussion on the limitations of democratic deliberation.

Recommendation 13: ***research funding agencies and other national and international agencies*** *should consider, as part of their toolkit, including forms of public deliberation as a means of heightening awareness and building legitimacy concerning the use of New Forms of Data in social science research. This could also include evaluation of these interventions; the building of an evidence base for public opinions on New Forms of Data and their use; and tracking opinion over time*

REPORT OF THE EXPERT GROUP

Introduction

Social science research using New Forms of Data has the capacity to yield improved insights into society, health, the economy and political behaviour. The creation, analysis and curation of these datasets gives rise to new manifestations of familiar problems: legal issues relating to privacy protection and intellectual property rights; technical problems concerning methods of analysis, storage and processing; and methodological problems relating to the quality of the data and its relationship to the population. Not least, these New Forms of Data create new ethical challenges.

The current investigation is being undertaken at a time when huge quantities of new and varied forms of data are becoming available due to dramatic changes in technologies and their application. While the future cannot be predicted with confidence, it is important to ensure that the safety and rights of those who are the subjects of research and analysis based on such data are upheld and that research is done in ways that also provides societal benefits. For this reason, we must return to basics and consider anew the values and principles that are at stake. The articulation of ethical responsibilities and requirements and the communication of potential public benefits should contribute significantly to a strengthening of reasoned trust between data subjects and researchers.

In the pre-digital 'old order' two situations predominated in the collection of personal information for research: people were either legally obliged to respond to enquiries (*e.g.*, a national census) or they were invited to contribute information and could refuse. The level of protection of access and thus the depth of analysis possible in these data sources depended either on a legal obligation or on the type of consent that the subject had given. In the evolving new order, a third situation has moved to the fore, where information about people is harnessed, often in conjunction with other data, in the context of providing a service or fulfilling other functions. Opting out of plans for research use of data collected in this manner (or even providing consent for different levels of re-use) may be possible in some cases but in others participation is mandatory. Although research will often not be the primary motivation for collecting such information, it may have considerable research potential. An important part of this changing landscape is New Forms of Data. They include a diverse range of data and data sources such as information on internet usage and commercial transactions; tracking data; and government information (see **Appendix 2**).

An important consequence of this evolution is that the life cycle of the single research project and the life cycle of the data intersect in different ways than in the past, leading to new questions about the respective roles of consent, information and security measures involving not only the researchers and data subjects, but also ethics review bodies, funders, policy makers, data services, and the public at large.

This report explores ethical issues related to the use of New Forms of Data for social science research and considers the steps that are needed to address these issues and ensure public trust in research. This 'ethical perspective' is complementary to recent work on this issue by OECD and other bodies, which focuses more on legal aspects of data privacy (see Box 1).

Box 1. International guidelines for research data management

This listing is not fully comprehensive and has a particular focus on recent OECD work. Other national and regional legal instruments designed to protect the privacy of individuals are discussed in detail in section 3 and Appendix 3.

OECD Principles and Guidelines for Access to Research Data from Public Funding 2007
These Principles and Guidelines advocate more open access to research data and are designed to assist all actors involved in trying to improve the international sharing of, and access to, this data.
http://www.oecd.org/sti/sci-tech/38500813.pdf

The OECD Privacy Framework 2013
These Guidelines constitute an update of the original 1980 version that served as the first internationally agreed upon set of privacy principles. Two themes run through the updated Guidelines:

- A focus on the practical implementation of privacy protection through an approach grounded in **risk management,** and
- The need to address the global dimension of privacy through improved interoperability.

http://www.oecd.org/internet/ieconomy/privacy-guidelines.htm

OECD Health Data Governance: Privacy, Monitoring and Research 2015
While focusing on the value of personal health records for research, this report emphasises the importance of strong data governance frameworks designed to assess the risks and benefits of the use of personal data in health research.

http://www.oecd.org/publications/health-data-governance-9789264244566-en.htm

Draft Recommendation of the Council on Good Statistical Practice, 2015
The OECD Committee on Statistics and Statistical Policy submitted draft recommendations on good statistical practice to OECD Council in September 2015. These recommendations have been developed for Members and Partners who wish to benchmark their statistical systems to OECD good practice.

http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=331

Ethical Protocols and Standards for Research in Social Sciences today. Science Europe 2015
This workshop document outlines the need for a more all-embracing framework for research ethics in the social sciences.

http://www.scienceeurope.org/uploads/PublicDocumentsAndSpeeches/SCsPublicDocs/20150911_Workshop%20Report_Social_Ethics_web.pdf

1. Ethical principles

Although current developments in information and communication technologies offer important and exciting opportunities for social science research, the new constellation of research agendas, technological capabilities, and analytical skills creates important ethical challenges. These concern a range of issues including: legitimacy, the impact of research, public involvement, consent, privacy and anonymity, social or economic discrimination, and distributions of power and responsibilities. These challenges are best grasped in the light of basic ethical principles. Such principles are not specific to research using New Forms of Data; they are spelled out here to provide a basis for the subsequent recommendations set out in this report.

The ethical principles relating to social research fall into two broad groups. Firstly, there are ethical principles that concern how researchers relate to each other and their own methodological standards. Here,

constructively critical openness is a crucial concept[3]. Secondly, there are ethical principles that concern how researchers relate to the surrounding world, primarily the data subjects. Here, the basic principles are *respect*, *ensuring good consequences*, and *justice*[4]. Both groups of principles are necessary in order to attain ethically sound and socially inclusive and useful research.

In order to function rationally and in a way that can ensure robust and valid results, there needs to be a high level of *constructively critical openness* among researchers. Among the major challenges in today's research environment is a tendency not to share data. At the same time there is growing concern about the validation of published scientific results, which is dependent on the accessibility of underlying data.

Good consequences are not easily defined in the abstract, although they (and their converse, bad consequences) are generally more recognisable when one considers a specific research project. In a time when the consequences of research are seen as inextricably entwined with other dimensions of social and political life, and the face of science is altering rapidly in conjunction with technological developments, it is especially important that this principle is reflected in research. Today, new uses for data are often suggested long after their initial collection. The implications of this are that long-term effects need to be addressed more systematically than is currently the case, and that recommendations and guidelines should deal with all elements of data handling rather than focusing only on the original collection of data.

Respect for persons has traditionally been demonstrated through the process of seeking voluntary informed consent. This is based on the notion that respecting persons includes respecting their right to make their own choices. For consent to be sufficiently informed, potential research participants need to be made aware of a number of issues, including how their privacy and confidential information would be protected[5]. In the world of New Forms of Data, such choices may not have been offered or made sufficiently clear. Divergence from this principle cannot be justified simply on the basis that the level of risk to individuals or to organisations is deemed low or non-existent. Safety (good consequences) is not the same as respect, and therefore mechanisms that ensure their safety may not necessarily fulfil the function of ensuring respect for persons' autonomy.

The principle of *justice* has a plethora of possible applications. For example, today's world of swiftly developing social media technologies and the new opportunities these offer for research using New Forms of Data increases the potential for an unjust balance of power between the single individual and massive, data-gathering institutions, be they academic institutions, private corporations, nation-states, or larger political entities.

Although these general principles work to reinforce each other, it can readily be seen how the two ethical perspectives broadly construed – values guiding how researchers relate to each other and values relevant to data subjects and populations – can also conflict. For instance, potential conflict exists for practices of data sharing seen in relation to the values of security and privacy; for the articulation of purpose seen in relation to demands of efficiency and efficaciousness; and for proportionality (and minimality) seen in relation to the current practice of default storage. This makes it all the more important that researchers, funders, policy makers, the media, and the public at large are engaged when it comes to ethical reflection on the use of New Forms of Data in the social sciences.

Conclusions

In considering these ethical principles in relation to research using personal data, the Expert Group identified a set of eight basic rules for the ethical collection, sharing, and research use of New Forms of Data, and would encourage their consideration by all stakeholders involved in the research process:

Validation, coordination, and quality control of data are required to guarantee rational advancement and social usefulness, as is a culture of open, constructive and rational criticism between researchers. At the same time, it is crucial that this happens in ways that do not threaten the personal integrity or dignity of data subjects or the population at large.

Basic rule 1. Mechanisms for the safe and responsible sharing of personal data, including mechanisms for the protection of privacy of data subjects as well as for public input and accountability, should be established and made public by data owners/controllers. Data should be shared as openly as is feasible within the relevant legal and ethical constraints.

There is a need to ensure that social science research respects persons and does not undermine people's free and informed agency.

Basic rule 2. The default position should be that personal data is not collected, processed or shared without informed consent. Ongoing efforts to update consent for new and unanticipated uses should be made where feasible.

A clear articulation of why and how a given research project is to be carried out is crucial both as part of what it means to respect data subjects (*e.g.*, as a precondition for consent) and in order to ensure that the data to be analysed are relevant for the proposed research and managed appropriately.

Basic rule 3. Articulation of purpose should be provided before a research project using personal data is carried out. In many instances, this will entail the development of transparent long-term plans and mechanisms for communicating any updates.

Basic rule 4. With a view to both the impact of the research and respect for data subjects, data quality should be considered to ensure that it is fit to fulfil the stated research purpose.

There is a need to ensure accountability and minimise the potential for unwanted consequences. These could arise from, for example, the onward transmission of data from one organisation to another leading to a lack of clarity about who bears the responsibility for ensuring that any subsequent research carried out is ethical and respects the rights of the individuals or organisations concerned.

Basic rule 5. Before a research or data collection project is undertaken, care should be taken to consider potential negative impacts, for individuals or groups, arising from the proposed project. Any potential negative consequences should be weighed against societal benefits, taking account of any mitigating actions to reduce the risk or impact of potential negative consequences.

Basic rule 6. Unambiguous distribution of responsibilities should be agreed in advance of any research-related data handling.

An important consideration is the potentially complex balance of power between, on the one hand, major public and private stakeholders and, on the other hand, the individuals whose data is being handled.

Basic rule 7. Data holders, research funders and researchers have a responsibility to consider how their role in the proposed research would contribute to the balance of power and influence between their institutions and individual data subjects.

Basic rule 8. Data holders and institutions that employ researchers should ensure they have access to an ethics review body (ERB) with the capability to review proposals to use New Forms of Data for research.

These eight basic rules provide the framework for an ethical approach to research using personal data. However, they do not provide a clear agenda with allocated responsibilities for a set of actions that will construct the framework. What follows is an elaboration of these rules and a set of recommendations designed specifically to facilitate their application.

2. The governance of research using New Forms of Data

> *'Research Governance may be defined as the broad range of regulations, principles and standards of good practice that exist to achieve, and continuously improve, research quality [nationally and internationally]"*[6].

Research governance covers issues such as 'protection of research participants, the safety and quality of research, privacy and confidentiality, financial probity, legal and regulatory matters, risk management and monitoring arrangements' (Australian National Health and Medical Research Council 2011: 1).

Legislation dealing with the protection of private information in many countries allows for exceptions when it comes to research (see **Section 3**). There is an implied expectation that researchers operate in a system that can be trusted to uphold values of respect for persons and their private information, and that research (including further analysis of data) is generally undertaken in pursuit of beneficial outcomes. Good research governance is a critical element of maintaining trust in research.

Broadly speaking, the governance of research takes place in a decentralised system that builds on shared values, notably honesty, trust, rigour, transparency and open communication, as well as care and respect for all participants in and subjects of research. Typical elements of the governance system to help ensure the integrity, ethics and quality of research (usually academic) include scientific peer review, ethics review, education and professional standards for researchers, publication and critical reflection, as well as professional and institutional oversight. Overarching regulations to help ensure consistent interpretation and application of standards also play an important part in the overall governance of research.

In some countries, there may be no formal set of legislated guidelines or associated oversight body or function available to researchers to advise on effective research governance. In countries where such guidance is available, it may not be equally available to all research disciplines. However, there are sufficient similarities across and within countries to allow generalisations about key stakeholders, shared values and common systems, and structures for implementing principles of good governance in research.

In most countries, before research is approved or funded it will normally be subjected to some kind of scientific peer review or quality control. This is an effective approach to ensuring that the research proposed is scientifically sound. Further to this, researchers or data analysts wishing to gain access to datasets containing personal data for secondary analysis may need to satisfy the data controller that they have successfully completed training on privacy and ethics, in order to be granted status of 'trusted researcher' – to meet the standard of 'safe people'[7]. The data controller may require proof that the proposed project or data analysis will meet ethical standards such as those relating to public benefit and respect for people / private or sensitive information. Hence, proof of ethics review of the proposed study may be required. The data controller and/or ethics review may require that the secondary data resulting from the proposed further analysis do not significantly increase the possibility of re-identifying individual participants and have been correctly analysed.

Commercial entities, including companies collecting and analysing large quantities of data that include personal information, are subject to principles of good governance and business ethics. Such companies are usually keen to avoid risks that might damage their reputation, or could potentially lead to claims for damages from individuals who are aggrieved about the use of private information for research or

other related purposes. Increasingly they are engaging in collaborations around New Forms of Data with public sector research partners.

New Forms of Data, including data generated from business activities and social media, have opened up opportunities for data collection and secondary analysis as well as experimental interventions to new actors who are not necessarily subject to the traditional governance systems for research. Appropriate governance of research that uses New Forms of Data is essential to promote and retain public trust in researchers and the research enterprise. **Table 1** below shows the major stakeholders and the general assumptions about their role in traditional research, together with the changes that may become more evident when research involving New Forms of Data is undertaken. **Table 2** proposes short-term and longer-term measures for facilitating good practice in research using New Forms of Data.

The Role of Institutional Review Boards/ Research Ethics Committees

Institutional Review Boards (IRBs) or Research Ethics Committees (RECs) are bodies established to independently review, advise on, and approve research involving human participants from an ethics perspective. These bodies (henceforth collectively referred to as Ethics Review Bodies or ERBs) are typically established at institutional or faculty level.

Many major funders of health-related research insist on independent review of research protocols by ERBs, and many countries have established national bodies to regulate or advise on ethics in research. In some instances, overarching legislation applies to all research involving human participants in the country, regardless of source of funding, as opposed to only research financed by funders setting such requirements.

Areas of oversight, composition, legal status and approach

Although there is no single set of internationally-recognised rules to regulate the composition, roles and functions of ERBs, it is possible to identify similarities. Particular focus areas for ERB oversight include:

- how participants will be selected and contacted;
- the way in which informed consent is to be obtained, how the privacy of human participants will be protected, and how any feedback will be provided on the outcomes of the research;
- how any risks of harm or unintended consequences affecting participants balance against the potential benefits to participants and the broader community (this may also be considered by the research funder);
- how potential risks can be mitigated before the research commences, or addressed if they materialise in the course of the research;
- whether the research proposes methodologies that are ethically sound and whether alternative methods exist that may reduce / solve possible ethical concerns with the proposed research.

ERBs are not, as a rule, separate legal entities. (One exception is commercial research ethics boards for health and clinical trial research in which multiple sites and commercial research houses are involved.) Institutional ERBs typically form part of the corporate governance structure of the institution where they were established. Levels of financial support and other forms of resourcing for ERBs are not standardised across institutions or across countries, and ERB members often render voluntary services with little or no

financial or in-kind compensation. Requirements for the size and composition of ERBs may differ from country to country.

Although ERBs may have powers to withhold or withdraw approval from research projects, they operate in an 'honour' system where the nature of the institutional support they receive to carry out their work and to safeguard their independence may co-determine whether they are able to provide optimal oversight.

ERBs may be required to play an oversight role in all phases of research, *e.g.* to review and approve (or exempt) protocols before research data are collected or analysed, to oversee compliance, respond to adverse events that may occur in the course of the research, to review and approve requests for amendments or continuation, and to receive reports following conclusion of the research.

In addition to institutional ERBs, a number of countries already have dedicated systems in place to manage access to large administrative datasets held by government departments or trusted data repositories. In the United Kingdom, for example, researchers requiring access to data containing sensitive or confidential information must first meet predetermined standards of training before they are granted the status of 'safe researcher' who may be granted access to such data under specific conditions. The question is whether the combination of data from different data repositories, not all of which may be carefully managed, could pose risks that fall beyond the area of expertise or jurisdiction of the stewards of such repositories. In cases where such risks are not sufficiently disclosed, known, or accounted for, ERBs at the institution where the researcher or lead researcher is housed could also play a role in reviewing the research plan to help identify and mitigate any potential risks.

Conclusions

Good research governance is critical to ensuring trust in research using New Forms of Data. Existing governance systems will need to evolve and adapt to the new data landscape.

Review of research proposals using New Forms of Data by an independent ERB is important to ensure that ethical issues and risks have been thought through and addressed before the research is undertaken. This review is important for each new research proposal. Independent ethical review will also help ensure transparency of the research process, and help retain and build public trust in the research.

It may be necessary to appoint or co-opt additional expert members to ERBs for the review of research protocols involving New Forms of Data, to consider:

- ethics issues relating to participants in research using New Forms of Data (such as recruitment, informed consent, how to deal with potential participants who are from at-risk populations, and disclosure);
- IT and internet-specific risks (*e.g.* data security and hacking);
- legal aspects of jurisdiction in multi-country research and cross-border data flows (including use of de-identified data only, and use for only authorised and consented purposes), and impact of existing rules of access to social media to which participants had already agreed.

Recommendation 1: ***all stakeholders with an interest in the use of New Forms of Data for research*** *should implement measures, based on current best practices, to strengthen the governance of research using New Forms of Data (see* ***Section 2, Tables 1 and 2).***

Recommendation 2: ***research funding agencies, research organisations and researchers*** *should ensure that any research proposing to use personal data is referred for review by an appropriate independent ethics review body.*

Box 2. The Facebook experiment and independent ethical review

Two university professors worked with Facebook Inc. to experiment with changes in the computer code that determines which social media posts from users' friends would appear on users' main pages. The changes created experimental groups that were slightly more likely to see posts containing words that express either positive or negative moods.

The research was approved by an internal ethics review at Facebook and later apparently presented as a request to analyse already existing secondary data, which must pass a lower ethical bar, to the Cornell Institutional Review Board. The legal authority for Facebook to conduct such research arises from its Terms of Service and Data Policy that users accept by ticking a box.

Research process and outcomes

Facebook selected nearly 700,000 users for the research. The subjects were not made aware that they were part of an experiment. The researchers were given access to all experimental subjects' de-identified postings in order to conduct automated measurement of the postings' mood. Analysis of the subsequent postings made by each group concluded that subjects' emotions were altered - they were mildly more likely to express the same mood that had been made more prevalent in the posts on their Facebook home page. After the research was made public there was a substantial public outcry and loss of trust.

Ethical issues raised

This research illustrates the problems of not making provision for a thorough and independent ethics review. An experienced ERB should be able to identify and consider issues relating to consent, the manipulation of mood, and the potential risks for adverse effects from the research or loss of trust by the data subjects and the public.

Looking more broadly, research has shown that emotions are correlated with propensity to vote in elections and candidate preferences. Mood manipulation could potentially be used in the future to influence voters' emotions and affect electoral outcomes.

Box 3. I don't need ethical approval…..do I?

A professor working in the linguistics department at National University is investigating the relationship between tweet messages and incidents of violence. The focus of the study is on tweets that contain words relevant to their topic of interest (*e.g.* hate speech, derogatory terms used when referring to foreigners).

Research process and outcomes

The research team assembled for this task monitors the frequency of use of such terms and whether there are spikes in the use of these terms when incidents of violence occur or are reported on. By analysing the gathered tweets, the team found that they were able to learn a lot about individual tweet authors and the messages they send. There were a few words and phrases in particular that they thought could be regarded as potentially inciting incidents of violence. The professor decided that they should investigate the matter further. It was agreed that the students would join the social groups that they have been monitoring to date, start emulating some of the messages sent, and observe the impact of these new messages.

The professor did not consider that the research needed consideration by an ethics review body (ERB) because:

- The research in question was self-funded, hence there was no need to comply with funder requirements.
- Their work was undertaken in the linguistics department and the university only had an ERB for health-related research. As there were no health-related issues raised by the research she proposed.
- The tweets that she and her team analysed were already in the public domain.

Ethical issues raised

Despite the professor's views on the requirement for independent ethics review, the lack of such review would mean that no consideration was given to the potential for harm to data subjects (especially resulting from the student interventions) or the risks of identification, any mitigating actions, and the balance between potential for harm and the potential gain for society of the proposed research. In this situation there was no identifiably appropriate ERB, and no clarity on the institution's expectations on ethics review for non-health related research. Where there is no funding requirement (and therefore no assessment by funders), the institution will generally be the only body with the authority to stipulate ethics review as a mandatory requirement. Finally, public engagement to ascertain the public's views might have provided helpful guidance on the acceptability of the research, but the publicity required could also lead to biased responses and therefore invalidation of the research results.

(Unlike the previous example, this is a fictitious case designed to demonstrate the need for independent ethical review of research proposals.)

Table 1. Role players in research governance, with special reference to quality control, ethics review and New Forms of Data

Role player	General assumptions in the context of 'traditional' research	New research paradigm and use of New Forms of Data
Researchers	Academically trained, usually to postgraduate level, may be members of professional bodies. Typically employed by, and accountable to, an academic or research organisation. Generally required to obtain independent ethics review from a recognised body. May be held personally liable for transgressions of codes of ethics or professional good practice. Varying levels of awareness about research ethics and protection of private information, but awareness should be high in the research leaders.	May be 'citizen scientists' May be computer programmers, or computer programs (if automated) May be self-employed May be employed in the private sector Much of their collection of personal data may not have been subject to any formal external ethics or quality review. Varying levels of awareness about research ethics and protection of private information. The position regarding liability may be unclear.
Research organisations	Usually academic institutions recognised as research organisations by research funders. Have systems and structures to support research administration and meet research compliance requirements.	May be for-profit business entities (including multinational companies), internet-based companies, not-for profit organisations. New actors may not have systems/structures to support research and compliance.
Ethics Review Bodies	Usually appointed at institutional, regional or national level In some instances, commercial ERBs are established to review research protocols from a particular sector, *e.g.* the pharmaceutical industry, to provide independent review of protocols for clinical trials.	Research units in for-profit organisations may have internal procedures to consider ethics of research in relation to any reputational risk to the organisation. Researchers or data analysts are unlikely to have much access to other ethics review.
Research funders	Often publically-funded grant-making institutions, or private foundations, usually with peer or expert review systems Conditions of grant often concern governance. May have policies or guidelines on best practice in research. These can be valuable resources to other countries or funders[8]. Consultancies, commissioned or contracted research are usually funded by government departments or commercial concerns. (May be proprietary research with no open access to findings or re-use of data.)	Research may be funded from own resources, crowdfunding, or funded by the institution employing the researcher. Review and quality control processes prior to funding decisions are not necessarily done by experts, peers or in relation to known criteria. In most instances, there are no explicit funding agreements of grant conditions in place. Issues around intellectual property, research ethics, governance, access to data, and re-use of data are often not covered in funding agreements (if there are any).
Research publishers	Formal peer review usually precedes decisions to publish or not to publish. Additional prerequisites for publication may include confirmation of ethics review, and access to the data used.	Research findings are often published on-line on sites where review and discussion follows publication.
Education and training institutions	Undergraduate curricula are usually discipline-bound and may not include training on ethics and privacy issues in research.	The advent of affordable or free Massive Open Online Courses (MOOCs) allow access to modular training programmes on ethics and privacy issues.

Role player	General assumptions in the context of 'traditional' research	New research paradigm and use of New Forms of Data
Professional bodies or learned societies	Professional bodies in the social sciences and humanities do not generally set standards for entry or continuous professional development.	New (often online) interest groups are emerging and contributing meaningfully to discourse on the ethics of New Forms of Data[9].
National advisory / policy bodies	In some countries there are excellent advisory bodies in the field of research ethics that publish guidelines on various aspects of research ethics[10].	
International bodies	Work of international bodies such as the OECD, WHO, UNESCO, European Commission help to identify issues, align standards and share best practices internationally.	
Custodians of data	Curation and wider sharing of research data is now encouraged. Limited access to administrative data (mostly census, sometimes education or health) may be managed by designated government officials. Business data are regarded as proprietary and generally not available for academic research. Exceptions are likely to be in response to individual requests; transparency and consistency in the reasons for providing access may not be evident.	In some countries, data repositories – especially for data generated from public funds – have ongoing management and curation, with guidelines to obtain access[11].
Other role players, including the media, and the general public[12]	Public outcry about perceived transgressions of human rights has led to changes in research governance, *e.g.* the publication of international codes, national guidelines, or legislation to govern research (especially health-related research) and the establishment of bodies such as ethics review boards. In some instances, 'lay persons' (*e.g.* representing the interests of the general public), or persons able to voice concerns on behalf of minorities or vulnerable groups, are required members of ethics review boards.	The media and general public may help to raise concerns around privacy issues or if research involving New Forms of Data is perceived to be unethical or improperly conducted. On-line resources similar to the Retraction Watch blog (which shares information on retracted journal articles and the reasons for retractions, to help increase transparency and accuracy in the scientific endeavour, see http://retractionwatch.com/) could potentially help raise privacy or other ethics concerns around research involving New Forms of Data.

Table 2. Measures to facilitate good practice in research using New Forms of Data

Role player /stakeholder	Measures that could be implemented quickly	Measures requiring more development
Researchers	Clarify the position regarding legal liability for breaches of data use policies.	Ethics training requirements for researchers / persons requesting access to personal data ('Safe people')
Research organisations	Review process used for safeguarding personal information and sensitive data, *e.g.* requirements to destroy, or submit for further curation and anonymisation, datasets that were created by combining original data.	Review systems and structures for the support of researchers and research governance. Consider implementing special arrangements to safeguard and protect data (*e.g.* data can only be accessed at specific physical sites; limitations on downloading of data) ['Safe settings'] Review process for assessing results following analysis to make sure individuals cannot be identified in new datasets resulting from analysis of combined data ['Safe outputs'].
Ethics Review Bodies (ERBs)	Establish publically accessible list of bodies providing ethical review of research using personal data (locations, type of research under consideration, proposal submission procedures etc.) Promote cooperation across borders between ERBs in the form of meetings to discuss developments, principles, topics and cases that affect both/all parties,	Relevant public bodies or regulatory agencies could accredit 'private sector' ERBs (as is the case in several countries already), and /or provide advice on controversial or emerging issues. Develop standards for recognising or accrediting ERBs with the capacity and experience to oversee research involving New Forms of Data – thus also opening the possibility of commercial or cross-border ERBs for studies involving New Forms of Data, or for ERBs with experience in reviewing research on New Forms of Data to assist ERBs with less experience. Requirement that ERBs consider applications before access rights are provided ('Safe projects')
Research funders	Where not already the case, issues around intellectual property, governance and ethics, access to data, and re-use of data could be explicitly covered in grant conditions before the research commences. Where funders have developed policies, guidelines or best practice notes for research that they fund, these could be made readily available as resources for others.	Fund training programmes for applicants, members of internal or external evaluating committees, or relevant Ethics Review Bodies
Research publishers	Ensure proof of ethics approval of studies before the results are published.	Establish guidelines concerning the data for publications: from ethical statement to citation of the data and access rights for others to (re-)analyse the data Involve watchdog organisations or individual critical reviewers to create more awareness of ethical use of New Forms of Data in research
Education and training institutions	Review undergraduate and postgraduate training to include training on ethics and privacy issues in relevant disciplines, including the fields of computer science, statistics and data analysis.	Encourage the development of affordable or free Massive Open Online Courses (MOOCs) and allow access to modular training programmes on ethics and privacy issues in research.

Role player /stakeholder	Measures that could be implemented quickly	Measures requiring more development
Professional bodies or learned societies	Establish special interest groups dealing with ethics and New Forms of Data to promote awareness or develop codes of good practice for researchers working in a particular discipline.	Researchers working in private companies could form or participate in special interest groups or engage with professional bodies to address issues around ethics and privacy concerns, and to help develop industry standards that are collectively agreed. Develop guidelines on good practice using New Forms of Data, with input from institutions and researchers.
National and international advisory / policy bodies	Ensure that all research involving personal data, not only research that has to comply with requirements of external funders, is subject to similar standards of ethics review and associated oversight.	Guidelines from advisory bodies dealing with ethics and New Forms of Data could be disseminated and discussed more widely, to allow review and adoption by institutions and further dissemination to researchers, research administrators and training institutions
Custodians of data		Custodians should ensure that, for data collected: • In the curation process, metadata could include information on the original 'informed consent' process, so it is clearly recorded what data subjects have agreed to. • Anonymisation / pseudonymisation for individual records; datasets only made available in aggregated form (with limitations in terms of analysis and combination of different datasets) • End user license agreements (EULAs) according to which persons accessing and using the data are expected to agree to specific conditions, *e.g.* correctly citing the data source and keeping private information confidential
Other role players, including the media, and the general public		On-line resources similar to the Retraction Watch blog (which shares information on retracted journal articles and the reasons for retractions, to help increase transparency and accuracy in the scientific endeavour, see http://retractionwatch.com/) could potentially help raise privacy or other ethics concerns around research involving New Forms of Data.

3. Legal frameworks

An ethical approach to the use of New Forms of Data for research purposes helps determine how such research should be undertaken, whereas the legal framework surrounding research data specifies what must or must not be done to comply with relevant laws. It goes without saying that research that does not comply with relevant laws should not be undertaken. An understanding of and compliance with these laws forms a crucial part of the wider ethical considerations. For this reason, an overview has been carried out of the main legal instruments relating to data protection that have been developed in the European Union (EU) and three illustrative countries (the UK, the USA and South Africa). This overview is presented in **Appendix 3** and has informed the analysis below.

Background

National and international legislation, directed at those who collect, hold, process and distribute personal data, has evolved gradually since the late 1870s. Initially focused upon 'the right to a private life', enshrined in the Universal Declaration of Human Rights adopted by the UN General Assembly in 1948, countries have been attempting to develop and adapt national legislation to cope with the rapid growth in electronic data, its sharing and re-use - including use for research. This is a complex area, with many countries developing their own legislation to govern the use of digital records of personal data, while others, particularly across Europe, are striving for a more harmonised approach. The situation remains in a state of flux, particularly so for the European Union as it moves to finalise and implement a new legal instrument for data protection and for the USA as it seeks to update the Common Rule.

Legal frameworks

From the review of selected regional and national legal frameworks governing access to personal information for research purposes (**Appendix 3**), it is clear that they have evolved partly in response to concerns about the need to promote research whilst protecting the privacy rights of individuals. Where research access to personal data is granted under conditions which could be regarded as an abrogation of these privacy rights, significant controls have been created in some countries to ensure that the balance between the public benefits arising from the research are weighed against any compromise of privacy rights. This 'public interest' test is particularly important where consent from data subjects has not been obtained for research use of their data. In such cases, or where the research is based on New Forms of Data that cross national boundaries (and therefore jurisdictions), ethics review bodies will normally be well placed to consider the specific risks and benefits.

Two related meta-questions that arise from this review of the legislative environments for research use of New Forms of Data are:

1. 'What impact do legal requirements have on the ability of researchers to conduct research using new and varied forms of data from persons or organisations?' For example, in the case of the UK, evidence collected as part of efforts to promote improved access to government-held personal data indicated that research was often not possible, or seriously delayed, usually due to a cautious approach to research access adopted by data controllers[13]. There is also concern about research based on New Forms of Data which cross national boundaries, where there is much uncertainty about the relevant jurisdiction. In the EU this will fall within the scope of the new Data Protection Regulation[14]. However, given that this has only recently been adopted, the new regulation has yet to be tested.

2. 'What are the ethical problems posed by legal frameworks which may, as yet, be unfit for purpose when applied to research using New Forms of Data?' This is a key area of concern, given that the limited evidence we have indicates that the legal frameworks lag far behind the rapidly developing technological capability to harness New Forms of Data for research.

Conclusions

These questions lead to the following recommendations:

***Recommendation 3**: **national and multi-national research funding agencies** should ensure that researchers have shown in their research plan that they:*

- *are cognisant of the relevant legal frameworks that may impact upon their access to and use of personal data for research;*
- *understand the adequacy of such legislation to protect the privacy of data subjects; and*
- *understand their legal responsibilities in relation to data collection, storage, processing, and sharing.*

***Recommendation 4: ethics review bodies** (ERBs) should ensure that their policy and practice can encompass the assessment of respect and privacy issues in proposals for data access and sharing where existing legal frameworks may not provide adequate protection for the data subjects, or where the data and/or research cross national boundaries. In the latter case, the ERB may need dialogue with the relevant researchers and/or ERBs in those countries. Recommendations made by the ERB in this respect should be incorporated into the research design and reflected in any subsequent use of the data collected*

4. Addressing key issues

4.1. Privacy

The notion of privacy has a role to play within each of the three basic ethical principles concerning people in research: respect for persons, due consideration of good/bad consequences, and justice. However, privacy is a concept that is not easily analysed. Privacy is often seen in terms of the capacity to control information about oneself. Whether we think of it primarily in terms of control or of access, it is important to keep in mind privacy's central role in enabling people to define, develop, and maintain their personal and social identities. Privacy is also central to the ability of an individual to participate fully in civil and economic life without retribution or discrimination. In what follows, we hold on to both the control and the access dimensions of privacy, while remaining sceptical about the possibility of providing a clear-cut definition of the concept.

What we mean when we speak of privacy tends to differ from one occasion to the next, partly depending on broader conceptual and historical contexts. Taking our cue from Nissenbaum (2011), we can say that there are at least three such contexts, each of which contributes to our initial understanding of the term 'privacy':

- *Private/public agency*. Throughout much of the last century, the notion of what is private has been shaped in contrast to what the modern state should be allowed to control. We thus speak of 'private citizens' and 'public agencies'.

- *Spheres of privacy*. These comprise a (probably open-ended) list of places and institutions. In this sense, the home, or a confidential conversation between close friends, are perceived as belonging to spheres of privacy, while an inauguration speech or the terminal hall of an airport are thought of as public.

- *Private information*. Privacy is crucially also conceived in terms of certain forms of information. Viewed through this lens, intimate facts about our sexual histories, or trivia such as whether one takes one's coffee with milk, can be regarded as private matters.

Many, if not most, disagreements on issues of privacy can be addressed by bearing this complexity in mind. However, analysis of privacy on the basis of these three contexts does not by itself amount to a heuristic for evaluating social science research using New Forms of Data - where the question of how to deal sensibly with the novel of the data context is at the heart of the matter.

Privacy controls

The new types of data being considered here pose considerable challenges to the standard principles surrounding privacy. It seems reasonably uncontentious to suggest that the privacy of a person (which includes organisations in the form of legal persons) must be safeguarded, and the means of that protection must be transparent.

Whilst individuals who use services that process their personal data should ideally have some choice over how their data might be used, it is not possible in practice to anticipate every possible case. A framework to protect subjects is needed to secure privacy rights, along with procedures to provide clear, intelligible information to subjects about the ways in which their privacy rights are being protected. Risks with the potential to curtail privacy rights of subjects need to be explained as well as the mitigating actions. These mitigating actions should also reflect any potential impact on specific groups.

The recently agreed international Framework for Responsible Sharing of Genomics and Health Related Data (Global Alliance for Genomics and Health, 2014), which advocates respect, the advancement of science, and a fair distribution of benefits and trust, provides a sensible starting point. The Framework states that a core element of good practice in responsible data sharing is for those involved to:

> *'Comply with applicable privacy and data protection regulations at every stage of data sharing, and be in a position to provide assurances to citizens that confidentiality and privacy are appropriately protected when data are collected, stored, processed, and exchanged. Privacy and data protection safeguards should be proportionate to the nature and use of the data, whether identifiable, coded or anonymised.'*

Conclusion

Due to the complexity of privacy, providing universal research ethics recommendations might risk either being too general to be helpful, or missing the mark as often as hitting it. In light of this, a new heuristic could provide a helpful approach for thinking through the topic of privacy in relation to specific research projects. A possible model is set out in **Appendix 4.**

Recommendation 5: ***Researchers and those involved in reviewing research proposals*** *should consider privacy protection, recognising that this is a complex issue with both legal and ethical aspects. For each proposal there should be a plan for clear communication to relevant audiences on how their privacy will be protected in research using personal data.*

4.2. Consent

At the core of what is deemed an ethical approach to research using personal data, and an approach that respects the right to privacy, is the concept of consent, where the individual whose data is collected is informed about the purpose of the research and consents to the use of their data for these purposes. In a research ethics context, 'consent' should be taken as short for 'voluntary informed consent' indicating that the subject has been provided with information on which they are able to make an informed and free decision about whether or not to consent. It is important to distinguish between the collection of data for

routine purposes and the use of this data for research. Consent may not be relevant to the data collection (e.g. tax data) but is likely to be required if that data is subsequently used for research. Data subjects should be made aware of the research purposes and the uses to which their data will be put, as well as any possible risks for them of the proposed research. In the light of this information, data subjects should be asked whether they voluntarily agree to the use of their personal information for research, decline its use, or place restrictions on the extent of use.

By obtaining informed consent from data subjects for the research use of their data, researchers may assume that they have surmounted a major ethical obstacle to their research plans. There are, however, a number of problems that relate not just to the process of obtaining informed consent, but more widely to the assumption that informed consent ethically validates the research process.

In operational terms, the process for obtaining informed consent includes providing a data subject with details of how and why their data will be collected and processed, and the steps that will be taken to protect the identity of the individual. Once obtained, the wording of the informed consent document sets the scope for use of the personal data; this must be respected if considering use of the data for purposes other than that for which they were collected. Operational issues may arise relating to: the nature of any consent already obtained; the capacity of researchers to convey the balance between possible personal risk and potential societal benefits accruing from the research; the time period for which consent remains valid; and the balance of power between the data controller and the data subjects.

An example of the manner in which these concerns are addressed can be found in Article 7 of the European Union Regulation on Data Protection[15], which states that:

1. The controller shall bear the burden of proof for the data subject's consent to the processing of their personal data for specified purposes.
2. If the data subject's consent is to be given in the context of a written declaration which also concerns another matter, the requirement to give consent must be presented distinguishable in its appearance from this other matter.
3. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal.
4. Consent shall not provide a legal basis for the processing, where there is a significant imbalance between the position of the data subject and the controller.

These requirements do little to address the fundamental issues identified by Barocas and Nissenbaum (2014). Researchers may not be able to anticipate the multiplicity and complexity of findings that could arise from research using New Forms of Data, nor does the information provided to data subjects or the public always address issues such as the data’s life span, or re-use of data for further research.

Informed consent and New Forms of Data

Typically, New Forms of Data used in research may have already been used in earlier research, whether for academic research or business analytics, and may be used again in future research. The composition and quality of these datasets may change, for example if additional data are collected, datasets are merged, or ongoing research alters the set as a whole at any given time.

Given that much of the power of analysing New Forms of Data may lie in the later re-use of the data in ways that the collector had not anticipated, issues around consent can be challenging, *e.g.* where no informed consent has been sought; where it is not viable to obtain informed consent; or where the informed

consent obtained is specific to the initial research project or business use for which it was collected and may therefore be invalid for different research using the same data.

Informed consent for the use of New Forms of Data for research, together with an undertaking to protect identity are viewed by many as the ethical foundations for the rapidly expanding research base using these New Forms of Data. However, the criticisms of operationally based aspects of informed consent and of the difficulty of data anonymisation for the protection of privacy, not just of data subjects but of all who could be affected by the research based on such data, have led some to claim that the ethical underpinnings provided by informed consent and anonymity are no longer sufficient in the era of New Forms of Data and Big Data. Similarly, protection against the identification of a data subject may require more than operational methods alone can provide.

Conclusions

For research using New Forms of Data, obtaining informed consent as a one-off event may not be the only appropriate way of ensuring respect for those involved as data subjects. Primarily, this is because it may make sense to think of consent in relation to the variety of foreseeable future uses to which the data could be put. In other words, the time may have come to consider 'following the data' and not only the projects. Moreover, the sheer complexity of data processing may force us to think anew about how data subjects can be kept realistically informed and empowered. It may be timely to consider new approaches to respecting the rights of those who provide personal data. One example of a new approach to consent for studies using New Forms of Data is the development of 'mobile consent', where consent is sought from large numbers of potential participants in a research study via a smartphone app[16]. This is a user-centred approach rather than a document-centred one and might offer data subjects low- or no-cost ways of staying informed and consenting to re-use of their data. Another approach (the two are not mutually exclusive) is the system for bank-like administration of individuals' personal data envisaged by Greenwood *et al.* (2014). This is an example of a technologically realistic way of ensuring dynamic and updated knowledge and control, on the part of the individual, concerning who gets to do what with their personal data. These and similar examples remind us that the technological developments that co-create the ethical challenges we now face also harbour possibilities for new ways in which to secure and improve informed consent and other ethical practices.

Better public engagement, informing all who stand to be affected about the possible risks and potential benefits from proposed research, will also help to address the limits of informed consent in the new data era. Above all else, as new approaches to ongoing communication and consent with research subjects are developed and tested, there is a need for a more harmonised approach to the regulation of procedures for consent; for the governance of research where informed consent is not an option; and for the maintenance of data protections designed to prevent disclosure of identities.

These considerations lead to the following recommendations:

***Recommendation 6: researchers** should:*

- for any research plan (whether it includes collecting new data or uses previously gathered data) produce and make available a brief statement understandable to non-experts, explaining the general purposes and motivations for the research, together with an assessment of the potential risks to individuals or groups associated with the data to be used for research;
- consider the means of obtaining, and wording of, the consent sought for new data collection with a view to 'future-proofing' the consent to enable future research projects to use the data and, where possible, offer the public and in particular research participants the means to receive updates about the progress of the research, including previously unanticipated uses of data and opportunities to reaffirm consent for use, where applicable.

Recommendation 7: ***data controllers, research funding agencies, ethics review bodies and researchers*** *should give careful consideration to the nature of any consent already obtained or required for the processing of personal data for research. Is it valid for the specified research? If not, can consent be obtained?*

Recommendation 8: ***ethics review bodies*** *should, where consent for research use of personal data is not deemed possible or would impact severely upon potential research findings, evaluate the potential risks and benefits of the proposed research. If the proposed project is deemed ethically and legally justified without obtaining consent, ethics review bodies should ensure that information is made publically available about the research and the reasons why consent is not deemed practicable, and should impose conditions that minimise the risk of disclosure of identities.*

4.3. Anonymity in the era of New Forms of Data

How well can the privacy of data subjects and respect for them as individuals be protected through anonymisation and de-identification? Before exploring this issue, it is worth emphasising that social and economic researchers dealing with massive amounts of data rarely need to know the identity of data subjects; their concern is with the relationship between variables within the data they seek to analyse. Direct identifiers, defined as explicit pieces of information in any dataset that yield knowledge about the identity of a data subject (*e.g.* name and address, date of birth), may be required for the purpose of creating a research dataset via linkage between different data sources, but are normally of no interest to researchers after linkage has been achieved and can be removed. Why then do we have a problem with anonymity and how do New Forms of Data exacerbate this problem? This section addresses these questions.

Anonymised data are data which have had identifying information permanently removed (and also perhaps been altered in other ways) to ensure that the risk of data subjects being identifiable is negligible.

For some research approaches to be feasible, a trusted party needs to hold identifying information in order to manage, curate, and link datasets for researchers. If the identity of an individual or an organisation cannot be inferred from a specific dataset, because direct identifiers have been replaced with an identifier which has no meaning to anyone but the data owner, the dataset is deemed to be 'de-identified'. This means that, in principle, there should be very little risk of the disclosure of identities provided that the data are made available without any further linkage or data matching being allowed.

Prior to the widespread introduction of electronic data files, matching between different paper records was seldom possible. The advent of electronic record-keeping and the widespread availability of detailed databases have changed this situation radically. Different datasets can now be easily matched and overlaid. Attempts to de-identify data simply by removing direct identifiers cannot therefore provide any confidence that the identity of a data subject is unlikely to be revealed[17].

Combining datasets can increase the risk of identification of individuals and/or sensitive information not readily available if each set is considered on its own. Given the rapid increase in both the number and size of datasets available, crucial questions in evaluating the safety of a given dataset are: what other information already in the public domain, and which other available datasets, could be combined with the first set and what risks could such combinations lead to, especially for data subjects? Determining the degree of access researchers are given to a dataset therefore needs to take into account the risks of disclosure inherent in that dataset and also the availability of other datasets that could increase those risks.

Can privacy be protected?

Various techniques have been and continue to be developed to prevent the disclosure of identities from within any given dataset[18]. Data protection 'by design and by default' is part of the EU General Data Protection Regulation[19]. The Regulation promotes techniques such as anonymisation (ensuring that the risk of somebody being identified in the data is negligible), pseudonymisation (replacing personally

identifiable material with artificial identifiers), and encryption (encoding messages so only those authorised can read it) to protect personal data. The purpose is to encourage the use of Big Data analytics, which can be done using anonymised or pseudonymised data.

Each of these techniques has its advocates, but the main problems relate to the expertise required to ensure that they have been correctly implemented and the resource implications. Many communities and/or countries therefore now take a more holistic approach - the use of a 'safe setting' or 'research data centre' for data access, matching and/or analysis.

What is a safe setting and how does it provide privacy?

The identity of data subjects cannot be protected simply by placing data in a location deemed 'safe'. The concept of a safe setting as a means of respecting the privacy rights of individuals requires mechanisms and security measures, which include the application of five strict (responsible and ethically robust) conditions. These five conditions are often termed the 'five safes':

1. Safe people: researchers need to be vetted for their ability to work appropriately with data.
2. Safe projects: research projects need to be approved within an ethical framework.
3. Safe data: risk of disclosure is minimised appropriately (if researchers do not need access to certain information, access to it should not be provided).
4. Safe environment: data can only be accessed in rigid, security assured locations.
5. Safe outputs: outputs of analysis must be checked for disclosure before they are released out of the safe environment.

These conditions minimise the risk of disclosure, but do not prevent it entirely. The heart of the matter is that, in a safe setting, disclosure is technically restricted to the researcher for a specific purpose and often within very specific legal constraints.

Implemented together these five safe conditions work towards minimising the risk of disclosure. Each of the five conditions can be applied at different levels. A safe setting may be a secure remote access mechanism, or a safe room. The correct level of application of one condition may depend on the level of application of each of the other conditions, or the appetite for risk of the data owner. We can say with some certainty that all of these conditions would be applicable to New Forms of Data, but experience with the operation of the processes to implement these conditions is currently limited in the case of such data. Clearly, further work on the important topic of safe settings and protecting privacy and identity needs to be undertaken.

Conclusion

From this analysis, two recommendations arise:

Recommendation 9: ***research funding agencies*** *should encourage further research on the development of statistical methods and software to provide assurances that the privacy of subjects in research using New Forms of Data is maintained, well understood and easy to implement.*

Recommendation 10: ***data holders, research funding agencies, and researchers*** *should share best practices in the creation and operation of safe settings, ensuring that restrictions on accessibility are minimised whilst maintaining data security.*

4.4. Ethical aspects of the commodification of data

The commercial sector is often the creator of New Forms of Data, including Big Data. Many of these datasets are compiled for analysis aimed at gaining insights useful for business, for example insights into consumer profiles or behaviour. The commodification of data referred to in this report relates to the buying and selling of personal data. Data subjects may have agreed to 'sell' (or give) personal information to a company in exchange for services. The company, in turn, may then capitalise on this personal information by using or re-selling it (or access to algorithms that were developed from the data) for purposes of further research or targeted advertising. The personal data thus becomes a commodity in the hands of users or traders. This commodification challenges each of the basic research ethics principles: justice, respect, societal benefit, and rationally critical openness. This may be further complicated by a lack of clarity about ownership of the data. It is unclear what the limits of acceptability should be where data access is limited for profit motives but a greater degree of openness could certainly, in some instances, enable research for public benefit.

Vomfell *et al.* (2015) have described the increased popularity, in recent years, of trading data through a new business model generated by emerging data marketplaces, sourcing data from a variety of sources before processing and trading them.

Benefits and challenges

Since New Forms of Data and information on the internet are often available for wider use and analysis by others, they can easily be exploited directly or indirectly by the market economy through large internet service providers who have the ability to build large databases by borrowing from various sources. While this is not a problem as such, issues arise when the use of information and research data by one agent is likely to preclude others from doing the same – often as a result of commodification. This is especially the case in the context of cloud computing, where the content holder has become as powerful, if not more powerful, than the copyright owner. Regardless of their legal status, information systems and large databases are increasingly controlled by large corporations who can control the manner in which they can be used or accessed.

There is increasing collaboration between the public and commercial sectors due to the potential for mutual benefit in ensuring the availability of high-quality data for scientifically and ethically sound research. There is scope for these partnerships to facilitate good practice on consent and privacy issues, and for the sharing of new methodological approaches. It is essential that high ethical and scientific standards are maintained by all parties involved in the collection and use of New Forms of Data in order to generate and retain public trust. Issues may also arise around ownership of data and other intellectual property generated through these partnerships.

Conclusion

There is a need to respect legitimate commercial interests whilst ensuring that the public can gain benefit from the research potential of the many and varied types of New Forms of Data held by private sector companies. While data holders, research organisations and researchers will negotiate their way through this complexity, there is a need to ensure that the public are fully aware of any for-profit transactions underpinning access to and the research use of private sector data.

Recommendation 11: ***For data holders, research organisations, and researchers:*** *Where personal data are bought or sold on a for-profit basis to inform research, information about the nature of these transactions should be included with reported research results.*

4.5. Data sharing, curation and security

The sharing of research data is used to inform new research and to enable the replication of existing research. In the past, data archives have collected discrete research datasets, prepared them for preservation and access, and made the resulting data collections available for re-use. This paradigm cannot be directly applied to all New Forms of Data for various reasons, including ethical requirements, some of which have already been considered earlier in **Section 4**.

This sub-section of the report is structured in two broad parts. The first provides a preliminary examination of the ramifications of New Forms of Data for the research data life cycle and data curation processes; the second examines existing procedures and controls which may continue to be applied when dealing with these data.

New Forms of Data and the data curation process

Until recently, most (but by no means all) data used for social science research was collected for the purpose of specific research. The 'traditional' principle that data should not be used for purposes for which it is not collected was in this respect reasonably easy to enforce, especially when the research has not been defined too narrowly. However, an important characteristic of New Forms of Data is that these are data which have research potential but were not necessarily collected for research, and may even not meet the data protection principle on fair processing.

Successful data sharing depends on successful data curation. Data curation has been defined as 'the activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purposes, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and other published materials.' (UK Data Service 2016) Data must be actively curated in order for access to be provided.

However, within many disciplines, including the social sciences, there is often a discontinuity in custody for data when it is transferred from the 'owner' to the 'archive'. Therefore, while this definition of curation retains currency in practice, it should be acknowledged that the curators can only influence the creators to a certain (sometimes very limited) extent. Most of the creators of these New Forms of Data are well beyond the traditional sphere of influence of the curators; thus this discontinuity can be magnified.

The traditional research data lifecycle has a long pedigree, and has been evolving over the last 50 years. The key features for this report are that the research design is undertaken *before* data collection (at least in the first instance), and that the requirement for re-use is understood to be part of that lifecycle. If data collected as part of this process are disclosive of identities of data subjects, the data owner can minimise the risk of disclosure by various techniques including anonymisation, and the repository can ensure that the data is stored securely (protecting against data loss and unauthorised access) and manage the risk of misuse through various methods at the point of access (see below).

What this report defines as New Forms of Data differ in their lifecycle from more traditional research data inasmuch as they are likely to have been created for a different purpose than the purpose(s) for which future researchers will use them. The data used in a research project thus comes with a history, the complexity of which can be further increased by that project.

The 'capture' of many New Forms of Data for research (which will often include data collected for business use) corresponds to the data collection phase in the traditional research data lifecycle; the key difference is that capture occurs in a context where ethical and legal control (and knowledge) of the data is beyond the influence of the repository and the researcher. This is not unusual - in traditional research a re-

user of data is in the same situation, but users are also supported through the mediation of controls provided by a digital repository but mandated by a data controller. These controls must continue to be supported within the 'new' model and should be applicable to all the spheres in which data use and data re-use can take place.

Controls and procedures

Curation activities (which include the provision of access to data) should minimise the risks for identification of subjects from data, and should maximise the opportunity for research.

The responsibilities of all third parties involved in maintaining data (including trusted third parties which may link different datasets) must also be clear, and the risks posed by any combination or linkage of data need to be taken into consideration.

Generically there are four key controls required for personal data to be successfully curated and thus amenable to re-use. These revolve around *consent*, *privacy rights*, the *maintenance of ownership*, (*i.e.*, copyright, intellectual property, *etc.*), and *research integrity*. These controls are usually maintained through procedures which deal with custody transfer (from owner to repository), data security (within the repository and as part of an access mechanism), and data curation (within the repository).

Most social science data archives implement these procedures through their own policy and governance structures, but different organisations implement them differently.

Any framework for the responsible use of New Forms of Data must explicitly cover the responsible sharing of those data beyond the original research team. In terms of the four key controls for successful data curation, (privacy, consent, ownership and research integrity), privacy and consent have been discussed in **Sections 4.1 – 4.3**, so will not be considered further here.

Ownership

For digital curators to stay within the law, the legal ownership of what they curate must be clear. The ability of a curator to undertake the actions needed to fulfil the definition of curation above depends on having a legal relationship with the data controller (usually the owner of the data). Creating and maintaining multiple copies of data may not be legal without this relationship being codified. Curating contextual information created by others than those who control the data, which allows the original data to be understood and used efficiently may require separate licensing, and thus ownership needs to be agreed. As data are combined, ownership is made more obscure by the application of intellectual property rights, so this aspect must be addressed and formalised by the responsible parties.

With New Forms of Data there is a need to face the difficult question about who owns personal data. If ownership cannot be appropriately assigned, then data sharing may be considered illegal even if the sharing can be deemed ethically responsible in other respects. The right for an individual to remove information about themselves from internet search engines under specific circumstances has been established; and while this fact does not imply that the person owns those data in a full sense, it does presuppose that the person has a level of control over their use. Curators of data, like internet search engines, therefore need to be able allow personally identifiable data to be removed at the request of the individual, and they must also manage the research integrity issues which may follow. In practice the former principle may be impossible to implement properly, *e.g.* where a data subject may know that he/she is included in a dataset, but the dataset has been de-identified before passing to a repository, and the repository cannot identify with certainty the individual in question.

Research integrity (and transparency)

Curation processes need to be transparent to maintain research integrity in the sense of research building systematically on, and not threatening, the integrity of data or results. The relationship between 'raw' data and 'data products' needs to be explicit.

Within traditional data archives, processes are usually in place to ensure that the content of any data file which is made available to a researcher has a clear provenance. Any changes made to a data file after it has been transferred to the repository should be documented, and researchers should expect that the data they access is identical to that any other researcher would have under similar circumstances (apart from the assigned and anonymised data identifiers). In order to maintain research integrity, in particular the transparency of research, a statement on research integrity should encompass the provenance and integrity of any data which is made available. It must also allow for the validation of research. The chain of responsibility for New Forms of Data makes the maintenance of reproducibility even more challenging. Similarly the development of protocols for allowing researchers access to data in order to legitimately verify the validity of each other's results, as part of building on each other's research, should be included in the development of broader curation processes.

Conclusion

For curation to be successful when applied to New Forms of Data, each of the four controls (consent, privacy, ownership, research integrity) needs to be applied, albeit in ways that may be different to those traditionally employed.

***Recommendation 12**: **researchers and data holders** should:*

- *establish guidelines and mechanisms through which applications to access data under their control may be made for the safe and reasonable sharing of data for publically-funded research;*
- *ensure that the requirements and processes for researchers to access the data under their control are made publically available;*
- *evaluate the potential for re-identification of individuals when depositing new putatively de-identified or putatively anonymised datasets, releasing datasets, adding new data, or developing the data access platform.*

Box 4. An ethical approach to cross-national data sharing

Maintaining an ethical approach to data sharing whilst protecting the intellectual property of those who have created research data collections can be challenging. An example of an approach is the *International Network for the Demographic Evaluation of Populations and Their Health* (INDEPTH), which has pioneered cross-national cooperation in health and population research. This is an interesting example of the way in which a group of, mainly developing, countries have come together to protect the intellectual property rights vested in data they have gathered, sharing data access and ideas on good practices, ensuring the curation and quality of their data whilst respecting the privacy of individuals and communities that provided the data.

The network consists of 53 health and demographic surveillance system (HDSS) field sites across Africa, Asia and Oceania, together with partner institutions and funders in Europe and North America. The network website details how interested people might make use of data collected at these sites. At the same time ethical principles are maintained via careful control over the prospective use of data.

The data sharing policy was established and is maintained by a Board set up for this purpose. It identifies various categories of data and access levels associated with each. It also stipulates the terms, conditions, scope and time frame for accessing and sharing the different data categories equitably, ethically and efficiently.

As an example, the Dikgale Demographic Surveillance (DDS) site, has been run and managed by the University of Limpopo (UL) in South Africa since 1996, collecting data on an annual basis. In a partnership between UL and the Flemish Consortium of Universities (VLIR-UOS in Belgium) which commenced in 2010, the DDS database at UL is being maintained and researchers in Belgium and UL who are able to utilise the DDS data for research purposes. This trans-national collaboration enables comparisons between INDEPTH surveillance sites.

Further information

http://www.indepth-network.org/data-stats/data-sharing-and-access-policies-and-protocols

Box 5. Data curation and the protection of identities

A PhD researcher has been thinking about conducting research on the relationship between dieting and obesity. She has signed on to an internet chat room with a fictitious name (as do all who join the chat room). The chat room is set up so that people with weight problems can chat openly and intimately with others about issues such as their perception of their body image, attempts to diet, *etc.* She realises that this is a source of useful information for her PhD.

She applies to her local ethics review body (ERB) for permission to undertake such analysis as part of her PhD. The ERB gives her permission subject to her ensuring that any information she publishes is suitably anonymised. The research is conducted via an analysis of textual 'conversations' that she has had with others in the chat room. The research is published in an anonymised format and the transcripts of conversations that she used are lodged in her institutional archive so that others may replicate her research.

One year later another student applies for access to these transcripts. On reading the transcripts, the student realises that the information almost certainly identifies one of the subjects as a well-known celebrity who has suffered from anorexia. The account is detailed and graphic. The student discusses this finding with a friend, who happens to be a journalist. An article then appears in a national newspaper entitled 'My near death struggle with anorexia'.

Ethical issues raised

This example illustrates the importance of the conditions imposed (and the adequacy of communication about these conditions and any expectations) by the owner of the chat room upon users regarding the behaviour of those who use its services. These could include *e.g.* validated registration of *nom-de-plumes* so that any misuse can be traced, and terms and conditions intended to prevent secondary analysis of the data. The example also illustrates the important role that could have been played by the ERB in considering how data would be made available for subsequent analysis, the potential risks for data subjects, and what institutional procedures would be needed to minimise these risks and help ensure the conditions imposed were not breached. Deception was used to collect the data, raising the issue of the extent (if any) to which deception can be justified. Finally, appropriate training would have taught the student that any discussion of these data with a friend would constitute a breach of confidentiality.

(This is a fictitious case designed to demonstrate the careful control of the conditions under which data may be reused for research.)

4.6. Public engagement

As stated previously, new findings from social science research based on New Forms of Data could have important and beneficial implications for the public, including research participants. However, there is a risk that any perceived invasion of individual privacy or other research activities which conflict with broadly accepted public norms could generate heated controversy and, with time, undermine the social

science research enterprise. As explained below, public engagement is essential for researchers to reach better ethical solutions on difficult issues and to ensure public acceptance of, and trust in, such research.[20]

Social science researchers using New Forms of Data confront difficult but important ethical questions concerning consent, privacy risks, and safeguards from harm. Such questions have no easy answers and poor solutions may greatly undermine public trust. Public engagement may be essential to identify ethical solutions acceptable to the public.

Directly communicating information about projects and their rationale to the public may reduce the risk of public distrust or outrage. Public input could also clarify how the reality or appearance of unethical activity could be avoided. The 'public' for pragmatic purposes could be the attentive public and opinion leaders that would be at the centre of any controversy. An example of public engagement to enhance awareness of the value and safeguards of new data social science research is the United Kingdom's Administrative Data Research Network. This includes a website that gives clear explanations of the importance of such research and the privacy and security precautions taken, a YouTube site with explanatory videos, a Twitter feed, and outreach efforts such as face-to-face public dialogues.

Beyond the pragmatics of public engagement to engender trust, the public is the legitimate arbiter of ethical issues, according to recent political and ethical theory. Well-known political philosophers and theorists have taken the position that only the input of an informed public can resolve difficult ethical issues affecting the public (Chambers 1996; Benhabib 1994; and many others). Habermas's theory of communicative rationality maintains that ethical conclusions are fully vindicated through widespread and thoughtful discussion of the issues (Habermas 1984). Chambers and Benhabib make informed public engagement critical to democratic legitimacy. If policy and ethical decisions will affect the public in important and widespread ways, then the public should decide such issues.

Formal Democratic Deliberation - Deliberative Polling

One useful format for formal democratic deliberation is the Deliberative Poll® (Luskin, Fishkin, and Jowell 2002). This is likely the most thoroughly studied deliberation format and illustrates a way to engage a public that is representative and that is made informed through the deliberative poll process. In a deliberative poll, as in representative surveys such as the Eurobarometer, a population of interest such as a country is identified, and is randomly sampled. The random sample of hundreds of people is invited to participate in a one or two day face-to-face or online deliberation, typically with a cash incentive. Participants are provided with balanced background information. The deliberation event consists of small group discussions with a trained facilitator and plenary sessions in which experts answer questions from the participants. A survey determining respondents' opinions and other variables is typically given before and after the event. Vulnerable and minority groups can be oversampled and given special opportunities to ensure input.

By bringing together a random sample of a population and allowing this sample to learn about and discuss an issue as citizens, deliberative polls seek to capture what the entire population would think about an issue if it were given the opportunity to learn about and deliberate an issue. Presumably, then, the public should find results from such polls particularly persuasive and legitimising.

Numerous other methods of public deliberation exist and can be utilised under different circumstances and constraints or for different objectives (see http://www.participedia.net/). Online deliberation can be used to cut costs, tele-democracy can reach a wide audience at low cost, citizens' juries can explore issues in great depth with a few participants, pyramidal democracy or sociocracy can permit meaningful interaction among large numbers of people, and online technologies and social media can highlight significant ideas or connect people. In countries or settings where access to such technologies for public

deliberation is limited, the inclusion of community representatives or lay persons in ethics review bodies or other engagement processes (including 'phone-in' services to local radio stations) could also be regarded as 'proxies' for consultation.

Research on Democratic Deliberation

Research on deliberation, typically in the deliberative polling vein, has repeatedly found robust evidence that such deliberation increases participant knowledge of the topic of deliberation and appreciably changes attitudes (Farrar *et al.* 2010). Other research finds that people are pleased with the outcomes of deliberation, saying they were satisfied with the experience and would participate again. More limited research suggests that participation increases the perceived legitimacy of institutions involved in the deliberation process, an important consideration for social science research using New Forms of Data.

Some research has shown limitations to formal deliberation, for example that women tend to participate less than men. It is not clear, however, how deliberation compares with alternative forms of public consultation, such as voting, which also exhibit demographic biases. In addition, research is needed on whether social psychological levers, such as the salience of the citizenship role, could be utilised to reduce demographic biases in deliberation contexts.

How best can the public be engaged?

The public knows little about new data collection, analytic techniques, or techniques ensuring data privacy. In addition, researchers have found a 'privacy paradox' in which the public states that it is deeply concerned about privacy but does little to protect its privacy online and has shown little willingness to pay to enhance such privacy. This is taken by some observers as evidence of confusion in the public. Can a public that is uninformed and conflicted provide, through deliberative engagement, meaningful input on new data social science research ethics questions and help legitimise decisions made?

A definitive answer can only be achieved by pursuing such engagement. Nevertheless, social science research ethics for New Forms of Data does not appear to be qualitatively different from other public policy issues that have been addressed with democratic deliberation, most of which involve low levels of public knowledge and, at times, confusion over objectives. The roots of the privacy paradox remain unclear. Part of the explanation may be the limited knowledge of the public with respect to how data is collected and what is done with it. Also, people appear to be highly insensitive to data collection and use that goes on outside immediate awareness and that affects the person in invisible ways. Two recent efforts to engage the public thoughtfully social science research issues related to New Forms of Data generally found that participants could learn about the issues and provide meaningful and coherent guidance (Patil *et al.* 2015; Cameron, Pope, and Clemence 2014).

Limitations

Democratic deliberation encompasses a range of methods that can be deployed for varying purposes and that have differing shortcomings. One basic limitation concerns the level of public interest in the issue under discussion. Interest will also affect the degree to which participants self-select into deliberations. Higher levels of self-selection affect how representative a deliberation's outcomes are for the general public. On the other hand, even a low response rate but with people representing a broad demographic range may provide insights about the public's informed views that would otherwise be left to guesswork. Moreover, democracies cannot expect everyone to be interested in every issue and the proper audience for a deliberation may be the portion of the public most capable of being interested in the issue. This may be the audience that could become activated in a controversy over the issue. Another limitation is cost, with representative sample, face-to-face deliberations being quite costly, though there are less costly alternatives.

Another limitation is that there is little systematic evidence that the results of a good deliberation can be used to sway the rest of the public. Much may depend on the success of publicising efforts. The persuasive effect of a deliberation might be enhanced by presenting the deliberation findings to members of the general public in a way that leverages the tendency of people to believe others who share their own values and identities - perhaps by showing that similar people in the deliberation came to the majority conclusion.

Conclusion

The following recommendation is proposed to improve the engagement of the public in research processes that make use of data that they may have supplied.

Recommendation 13: ***research funding agencies and other national and international agencies*** *should consider, as part of their toolkit, including forms of public deliberation as a means of heightening awareness and building legitimacy concerning the use of New Forms of Data in social science research. This could also include evaluation of these interventions; the building of an evidence base for public opinions on New Forms of Data and their use; and tracking opinion over time.*

5. Summary

In this report we have set out some basic rules that underpin an ethical approach to research using New Forms of Data. These rules and the interpretation that we place upon them give rise to a set of recommendations designed to provide a framework for the ethical governance of research using such data. There are assumptions and limitations underpinning these recommendations – they are not cost-free and will be easier to apply in countries with established research ethics procedures, particularly where research organisations and data owners have access to ethical review bodies. The sharing of expertise on and knowledge about research ethics between countries is critical to the creation of a common and cost-efficient ethical environment for social scientific research.

Some readers may view the recommendations presented here as creating obstacles, inhibiting research based on New Forms of Data. We argue that there is a careful balance to be achieved between the need to have in place ethical guidelines whilst promoting the research value of New Forms of Data. These recommendations have been formulated carefully and with a sharp eye on the direction that social science is moving. The abundance of data of different types with the potential to inform valuable research on our social and economic wellbeing exerts a strong influence on this direction – and it is here that risks may lie. However, we do not have perfect foresight and we cannot predict the ways in which data and research will evolve over the next few years. While new developments may help reduce these risks, there are clear principles that we consider as enduring. Steps can be taken to apply these principles. To have an awareness of the risks associated with the research use of New Forms of Data is an important first step. To establish systems and put in place structures that evaluate potential risks and weigh these against benefits is the next step. These may take different forms, depending upon the extent to which a framework for research governance in the social sciences already exists in particular countries. The recommendations in this report are intended to be useful for all those involved in social science research, whether as researchers, reviewers, funders, data controllers/holders, publishers, or policy makers.

Social scientists conduct their research in an atmosphere of trust, and trust will be eroded if there is a perceived misuse of personal data by some within the research community. An overarching aim for the recommendations presented in this report is to uphold this trust relationship between social scientists and the public.

GLOSSARY

These definitions were adopted by the Expert Group in order to ensure consistency of understanding across the Group. The use of one source as opposed to another does not confer additional legitimacy on that source.

Anonymisation

Anonymisation refers to a process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves more than simple de-identification (qv) but also requires that data be altered or masked in some way in order to prevent statistical linkage.

Source: UK Anonymisation Network.

Big Data (and New Forms of Data)

What is currently often called ‘Big Data’ is best grasped through considering three interrelated dimensions: size, approach, and technology. As for size, Big Data is big in a purely quantitative sense. But size alone is not sufficient to identify Big Data. The approaches, methodologies, and mind sets developed to handle such data sets are equally central – a feature which stems partly (but not exclusively) from the fact that much Big Data is co-defined by the data being unmanageable for research through direct human perusal due to size and/or variety. This again implies that the phenomenon of Big Data encompasses relatively new and rapidly developing technologies. ‘Big Data’ can be very large, complex and constantly changing and they may require new analytical techniques to maximise their research potential.

The focus on this report is on 'New Forms of Data' for social science research, which includes some forms of Big Data that are associated with individuals; this may include corporations. However, New Forms of Data as defined for this report is broader than Big Data and includes data, such as blog posts, that are by no means big, but are available for qualitative analysis in social science research. **Appendix 2** gives examples of New Forms of Data that have potential for social science research.

Consent

Informed consent in a research context entails giving sufficient information about planned research for prospective participants to make an informed and free decision on their possible involvement, without explicit or implicit coercion to participate. Information should be provided in a format that is comprehensible and accessible to prospective participants, typically in written form (or in a form that participants can access after the end of the research interaction), and time should be allowed for the participants to consider their choices and to discuss their decision with others if appropriate. Written consent forms should be signed by the research participants to indicate consent.

Based on: UK ESRC Framework for Research Ethics (2015)

(Digital) Curation

When the term curation is used in the text, it means all those activities which allow access to data in the longer term. A more formal definition is 'Digital curation is all about maintaining and adding value to a trusted body of digital information for future and current use; specifically, the active management and appraisal of data over the entire life cycle. Digital curation builds upon the underlying concepts of digital preservation whilst emphasising opportunities for added value and knowledge through annotation and continuing resource management. Preservation is a curation activity, although both are concerned with managing digital resources with no significant (or only controlled) changes over time.'

Source: ICPSR Glossary.

Data access

Data access is the activity by which a researcher is given access to data. The type of access depends on the known risks to disclosure (qv) of personal information.

Data controller

Data controller means a party who, according to domestic law, is competent to decide about the contents and use of personal data regardless of whether or not such data are collected, stored, processed or disseminated by that party or by an agent on its behalf.

Source: OECD (2013), *The Privacy Framework*, p.13.

Data holder

Any organisation or person holding a dataset that includes data about individuals (whether people or organisations).

Data owner

A data owner should generally be synonymous with a data controller (qv); however we suggest that a data owner should explicitly additionally include the 'ownership' of intellectual property rights as well as other rights defined. Thus, using the Open Digital Rights Language (ODRL) terminology, a data owner is a party (legal entity) with a role in defining permissions, duties and prohibitions including the approved actions or relevant constraints on actions in relation to a digital asset. Data owners have primary legal responsibilities, and are not the same as data processors.

Source: based on the Open Digital Rights Language terminology.

Data protection

Data protection is the process of safeguarding important information from corruption and/or loss, and includes legal control over access to and use of data stored in computers. In the context of personal data, data protection refers to the set of privacy-motivated laws, policies and procedures that aim to minimise intrusion into respondents' privacy caused by the collection, storage and dissemination of personal data.

Based on: OECD (2015) Glossary of Statistical Terms.

Data subject

The person or persons described by the data in a given dataset or collection of data.

De-identification

De-identification is the process of removing or masking direct identifiers from personal data. De-identification is not the same as anonymisation (qv), it is the process which removes those elements of personal data which can immediately identify a person (individual or corporate) from personal data. De-identification includes what is often called pseudonymisation.

Source: Modified from OECD Expert Group for International Collaboration on Microdata Access: Final Report and the UK Anonymisation Network document.

Disclosure

In the context of this report, disclosure relates to the inappropriate or unintended attribution of information to a data subject, whether an individual or an organisation. Disclosure has two components: identification and attribution. The authors of this report believe that one strategy to reduce disclosure risk is disclosure control (qv) which is the reduction (and not 100% prevention) of disclosure risk. Hence the unintentional attribution of information is explicitly included in the definition as part of disclosure.

Source: Based on OECD (2014) and ESSNet SDC (2010).

Ethical review

Review of the ethical implications of a project and how these issues are addressed. The main aim of an ethics review is, as far as possible, to protect all groups involved in research: participants, institutions, funders and researchers, throughout the lifetime of the research and into the dissemination process.

Source: ESRC Framework for Research Ethics.

IRB/REC

There are several definitions of 'Institutional Review Board' (IRB) or 'Research Ethics Committee' (REC). For the purposes of this report, they refer to a body (which may or may not have a committee-like structure), usually established by a research organisation, to review and approve research which involves human participants to ensure that their dignity, rights and welfare are protected and that the research which they approve is conducted in accordance with all relevant ethical guidelines. The rules according to which these bodies are constituted and operate are often developed using international guidelines and national legislation. Although the detailed responsibilities and the scope of the oversight of such bodies may differ between countries, especially in the domain of non-medical research, the way in which they are governed show many similarities. These bodies should have strict rules relating to conflicts of interest, and include members who have the necessary knowledge, expertise and experience to help promote good research from the perspectives of scientific excellence, good outcomes and respect for human participants, and have regular monitoring of and accountability for their decisions.

Source: Adapted from the *ESRC Framework for Research Ethics (2015)* and the American Public University System. [http://www.apus.edu/community-scholars/institutional-review-board/].

Metadata

Metadata provides information on data and the processes of producing and using data. Metadata are data which are needed for proper reproduction and use of the data.

Source: UNECE/UNSC (1995).

Personal Data

There are many definitions in use relating to personal data. For this report we understand it to be: any information relating to an identified or identifiable natural person ('data subject'). An identifiable person is one who can be identified, directly or indirectly. Where an individual is not identifiable, data are said to be anonymous.

Source: OECD (2015) Glossary of Statistical Terms.

Privacy

For the purposes of this report, we understand privacy as a concept which applies to data subjects while the concept of confidentiality applies to data. Thus we adhere to the UNECE *Principles* and use: 'Someone's right to keep their personal matters and relationships secret, involving an obligation of the holder of information to the subject of the information to do so.'

Source: UNECE (2009).

Reproducibility

Reproducibility is the ability for a piece of data analysis to be reproduced, either by the researcher or by someone else working independently, using the original data. Replicability by contrast is used in research integrity to mean the replication of a piece of research without using the original data.

Research

For the purposes of this report, we define research as 'a process of investigation leading to new insights, effectively shared. It includes work of direct relevance to the needs of commerce, industry, and to the public and voluntary sectors; scholarship; the invention and generation of ideas, images, performances, artefacts including design, where these lead to new or substantially improved insights; and the use of existing knowledge in experimental development to produce new or substantially improved materials, devices, products and processes, including design and construction'. The type of organisation in which research takes place does not alter the need for ethical review.

Source: UUK (2013).

Research data

Research data are defined for the purpose of this document as information relevant to, or of interest to, researchers, either as inputs into or outputs from research. They are research materials resulting from primary data collection or generation, or derived from existing sources intended to be analysed in the course of a research project.

Source: ESRC Research Data Policy.

Research integrity

Research integrity essentially refers to the responsible conduct of research and the trustworthiness of the research endeavour. Key principles underpinning this are *honesty*, *rigour*, *transparency and open communication* and *care and respect* for all participants in and subjects of research. In the context of data

curation, research integrity refers primarily to mechanisms and practices that ensure the integrity of the data used for research.

Source: UUK (2013).

Researcher

'Researcher' as used in this report may refer to someone carrying out research who is based in academia, a government or policy institution, or a commercial organisation.

Re-use

Traditionally, data re-use refers to the possibility of datasets being used for purposes for which they were not originally designed. While re-use is generally for new research it may also include the reproduction of previous research.

Safe setting

See secure environment.

Secure environment

A secure environment is a physical location and/or remote access infrastructures with procedures and protocols which ensure an appropriate level of control on data access, data use (execution) and release. This includes but is not limited to Secure Server Room, Secure Room, Safe Pod, Secure Remote Access solutions and Remote Execution.

Source: Adapted from Administrative Data Research Network Glossary.

Validation

A continuous monitoring of the process of compilation and of the results of this process.

Source: OECD (2015) Glossary of Statistical Terms.

APPENDIX 1. EXPERT GROUP TERMS OF REFERENCE AND MEMBERSHIP

Terms of reference of the Expert Group:

- conduct a review of activities, practice and problems across a range of countries regarding the ethical use of new forms of data for research;
- identify and analyse ethical issues that may arise from research use of various new forms of data, with particular reference to the balance between the social value of research utilising such data and the protection of the well-being and rights, including privacy rights, of individuals;
- convene an international conference to identify and promote best practice regarding the ethics of using new forms of data to address research issues in the social sciences and at the boundary with other scientific disciplines;
- draft internationally acceptable guidelines, suitable for adoption by research funding agencies, for an ethical approach to the use of new forms of data for research.

Membership

Australia	Gemma van Halderen
Spain	Txetxu Ausin
European Commission	Agni Kortsidaki Maurizio Salvi
Israel	Hagit Schwimmer
France	Pascal Buleon
Germany	Stefan Bender
Switzerland	Mike Martin
Japan	Ichiro Satoh Osamu Sudoh
Norway	Hallvard Fossheim *(Vice Chair)*
South Africa	'Maseka Lesaoana Christa van Zyl
United Kingdom	Matthew Woollard Peter Elias *(Chair)* Samantha McGregor
United States	Peter Muhlberger Charlie Catlett Kate Cagney Robert Goerge
Invited expert	Charles 'Chuck' Humphrey
OECD Secretariat	Carthage Smith Keiko Kimura Christian Reimsbach-Kounatze

APPENDIX 2. EXAMPLES OF NEW FORMS OF DATA

Broad category of data	Detailed categories	Examples
Category A: Government transactions	Individual tax records	Income tax; tax credits
	Corporate tax records	Corporation tax; sales; tax, value added tax
	Property tax records	Tax on sales of property; tax on value of property
	Social security payments	State pensions; hardship payments: unemployment benefits; child benefits
	Import/export records	Border control records; import/export licensing records
Category B: Government and other registration records	Housing and land use registers	Registers of ownership
	Educational registers	School inspections; pupil results
	Criminal justice registers	Police records; court records
	Social security registers	Registers of eligible persons
	Electoral registers	Voter registration records
	Employment registers	Employer census records: registers of persons joining/leaving employment
	Population registers	Births; marriages; civil unions; deaths; immigration/emigration records; census records
	Health system registers	Personal medical records; hospital records
	Vehicle/driver registers	Driver licence registers; vehicle licence registers
	Membership registers	Political parties; charities; clubs
Category C: Commercial transactions	**Store cards**	Supermarket loyalty cards
	Customer accounts	Utilities; financial institutions; mobile phone usage
	Other customer records	Product purchases; service agreements
Category D: Internet usage	**Search terms**	Google; Bing; Yahoo search activity
	Website interactions	Visit statistics; user generated content
	Downloads	Music; films; TV
	Social networks	Facebook; Twitter; LinkedIn
	Blogs; news sites	Reddit
Category E: Tracking data	**CCTV images**	Security/safety camera recordings
	Traffic sensors	Vehicle tracking records; vehicle movement records
	Mobile phone locations: GPS data	
Category F: Satellite and aerial imagery	**Visible light spectrum**	Google Earth©
	Night-time visible radiation	Landsat
	Infrared; radar mapping	

APPENDIX 3. LEGAL FRAMEWORKS IN THE EU AND THREE COUNTRIES

The situation in the EU

In 1978 the OECD established a group of experts, tasked to elaborate a set of principles governing the protection of personal data. Working in close collaboration with the Council of Europe (1950), seven basic principles of data protection were defined. These were:

- There should be limits to the collection of personal data, which should be collected by fair and lawful means and, where possible, with the consent of the data subject.
- Personal data should be relevant to the purpose for which they are required, should be accurate, complete and up-to-date.
- The purpose for which personal data are required should be specified not later than at the time of collection.
- Personal data should not be disclosed or used for purposes other than that for which they were collected, except with the consent of data subjects.
- Personal data should be protected by reasonable security safeguards against unauthorised access, loss, destruction, modification or disclosure.
- Means should be established to facilitate the existence and nature of personal data and the identity and residence of the data controller.
- Data subjects should have the right to gain access to their data, to challenge such data, to request erasure and to have the right to challenge any denial of these rights.

By 1980 the Council of Europe had proposed a Convention for the Protection of Individuals with regard to the Automatic Processing of Personal Data. This convention reflected the seven basic principles agreed by OECD member countries.

Throughout the 1980s, at the time of the EEC (European Economic Community, since 1993 the EU), the European Commission took inspiration from some principles of data protection expressed by the OECD and the Council of Europe. By 1985 the Council of Europe's Convention came into effect, but its adoption among the EEC signatory countries was uneven. Recognising this, the EEC published a draft data protection directive in 1990. After the necessary consultation period within and across member states of the EEC, the EU Data Protection Directive was passed in 1995; being a directive it proved helpful in providing a first comprehensive framework for the EU, although it required implementation at the national level.

The 1995 directive has been influential in shaping the legislation in EU countries. As new countries have been admitted to the EU, they are required to ensure that the aims of the 1995 directive are met. The

directive additionally gives citizens the right to access their personal data and to request it to be removed from processing if incomplete or inaccurate. Despite its influence, the European Commission concluded that, given the immense technological changes that had occurred since it was passed, the directive should be reformed and its operation strengthened by adopting a new piece of EU legislation.

The new EU General Data Protection Regulation (GDPR) is EEA-relevant and will be applied from May 2018 in all EU member states, plus Iceland, Norway and Lichtenstein. Major changes from the directive include:

- a single set of rules on data protection, allowing for a higher level of simplification and legal certainty, while reducing administrative costs;
- sensibly stronger protection of citizens' personal data while ensuring the application of suitable conditions for EU research;
- the need for unambiguous consent by data subjects. Processing for scientific research purposes may alternatively be justified based on public interest;
- the requirement for companies and organisations to promptly notify authorities in cases of serious data breaches to avoid very high fines;
- the 'right to be forgotten' (personal data must be removed from use if consent is withdrawn);
- the scope of the legislation – EU rules must apply if personal data is handled abroad by companies that are active in the EU market and offer their services to EU citizens;
- the establishment of national data protection authorities (DPAs) to be coordinated by a European Data Protection Board;
- the need for consent (valid consent for data to be collected and the purposes for which it will be used must be explicit rather than implicit).

The new legislation may have a significant impact upon the conduct of research using New Forms of Data and Big Data, whether generated by public or private sector bodies. While much detail remains to be resolved, it is likely that data processed for research purposes will be subject to more consistent scrutiny by data protection authorities than has hitherto been the case.

The situation in the UK

The broad protection afforded to personal data via legislation is covered by the transposition of the EU 1995 Data Protection Directive into UK law, creating the 1998 Data Protection Act. It requires that anyone who processes personal data must make sure that personal data are:

- fairly and lawfully processed;
- processed for limited purposes;
- adequate, relevant and not excessive;
- accurate and up to date;
- not kept for longer than is necessary;

- processed in line with your rights;
- secure; and
- not transferred to other countries without adequate protection.

While this governs the creation, handling, storage and re-use of personal data, it does not provide a clear 'gateway' through which a researcher may request access to specific types of data. However, there are two additional legal frameworks which govern access to and research use of personal data, one covering what are deemed 'official statistics' (basically statistical information identified as such by the national statistical authorities) and the other relates to health and social care data. In each case the scope of the legislation varies by country, so this overview focusses specifically on the prevailing situation in England.

The Statistics and Registration Service Act 2007 provides a mechanism whereby access to personal information in what are classed as 'official statistics' (statistical information produced by government departments and agencies) can be supplied to researchers under specific conditions. These are that the researcher must be approved by the statistical authority and that the researcher has signed a declaration indicating their awareness of the conditions under which access to personal information has been granted. The statistical authority is required to ensure that the person to whom such approval has been granted is 'a fit and proper person' and that the authority has considered the purpose for which access has been requested. There is no mention of ethical review forming a part of these consideration, but recently a new body has been established by the statistical authority that provides for ethical oversight in situations where there is no such oversight provided by the institution employing the researcher.

The National Health Service Act 2006 provides a legal 'gateway' that enables the common law duty of confidentiality to be temporarily lifted so that confidential patient information can be transferred to an applicant without the discloser being in breach of the common law duty of confidentiality. For this to happen, a number of requirements must be satisfied, notably the following:

- the activity for which personal data have been requested must be a 'medical purpose'. Medical purposes include medical research that has received ethical approval by a research ethics committee;
- the activity must be in the public interest or in the interests of improving patient care;
- the activity must be compliant with the provisions of the Data Protection Act 1998;
- all applications must undergo an annual review to evidence whether support is still necessary; and
- the person(s) receiving the information has undergone an independent review of their purposes and governance arrangements

Despite these two legal mechanisms providing access to and linkage between New Forms of Data, there are numerous obstacles facing researchers who are planning to gain access to various types of personal information for research purposes. These relate to data which fall outside the scope of the National Health Service or are not under the control of the statistical authority, including social welfare records, tax records, and licensing information of various types. While such data are covered by the 1998 Data Protection Act, public sector agencies either have specific legislation which may inhibit data sharing, or there is a common law duty of confidentiality which limits access. The introduction of new data sharing powers is currently under review, a process which could result in legislation to provide improved access to

government administrative data. There remains a notable absence of specific legislation covering many of the New Forms of Data that are the focus of this report.

The situation in the USA

The USA has not taken an overarching approach to the implementation of privacy legislation covering personal information of all types. A more sector-specific approach has been adopted, which applies with some exceptions to personal information held by federal agencies and specific regulated entities and relies to a significant extent upon separate legal instruments specific to the release of personal information contained in particular data sources (*e.g.*, Title 13 USC for Census data, Title 12 USC for financial .information, the HIPAA Privacy Rule for health information). The framework regulating the inclusion of human subjects in research is based upon the Belmont Report (US Department of Health, Education and Welfare 1978) and governed by the Federal Policy for the Protection of Human Subjects, also known as the 'Common Rule'. The first is largely a discussion of some of the ethical principles referred to earlier in **Section 1** of this report: beneficence, respect for persons, and justice. The second is a regulation that elaborates administrative procedures for protection of human subjects. Both of these documents were written primarily with biomedical and behavioural research in mind. Only the Common Rule is legally enforceable and applies solely to research conducted by certain federal agencies or undertaken with federal government funding. Ethical concepts are mentioned in the Common Rule, but are not the main focus of the Rule. Detailed ethical judgments are largely left to individual institutions and their institutional review boards (IRBs). The Common Rule requires IRBs to focus on the risks and benefits of the research to the research participants, and not to consider the risks and benefits in how any intervention or other outcome arising from the research might eventually be applied.

From the point of view of social science research using New Forms of Data, the existing U.S. legal framework may present some challenges. Because the framework was designed to protect subjects in biomedical rather than social research, informed consent requirements have been criticised by some as elaborate and time-intensive. Such requirements may be inappropriate or impractical for social science research, particularly that which makes use of internet-generated data. Because enforcement operates as a requirement through federal agencies or federal funding, research based on New Forms of Data and funded entirely with private sector money, whether conducted by individual companies or in conjunction with academic researchers, faces no required oversight via the Common Rule. Existing regulations also do not address 'privacy-related' ethical issues in which, for example, information collected or analyses done on a sample could be used to harm people outside the sample such as through price or employment discrimination (although IRBs do consider the concept of familial, or group harm in their review). Another privacy-related risk is that, though a dataset may hold no data which directly identifies individuals (*e.g.*, names, addresses, dates of birth) the 'data pattern' may be unique to each subject and permit linking the dataset with others, potentially revealing private information and allowing discrimination, with or without re-identification.

Revisions of the Common Rule are currently being considered by the U.S. Dept. of Health and Human Services and the other agencies that are subject to its requirements. Under the proposed revisions, oversight and informed consent procedures would be streamlined for research that poses little risk to subjects. As with the Common Rule, the proposed revisions apply legal oversight only to federal agencies and to research receiving federal funding. Provisions are not made for consideration of the broad societal implications of research, particularly how research might be used. The content of the revisions, however, remains in flux.

The situation in South Africa

Following the country's first democratic elections in 1994, a new Constitution was adopted in 1996. Enshrined in Chapter 2 of this Constitution is a Bill of Rights which represents many of the ideals that had been fought for in the struggle against Apartheid. The Bill of Rights provides an enabling framework for research and protects the rights of research participants. Specific clauses that are relevant to research access to information and protection of personal information include Section 12 (of the constitution), which includes the right '*not to be subjected to medical or scientific experiments without their informed consent*', Section 14, which deals with privacy and includes the right '*not to have the privacy of their communications infringed*' and Section 16, which deals with freedom of expression which includes '*freedom to receive or impart information or ideas*' as well as '*academic freedom and freedom of scientific research*'.

Beyond, and building on, the ideals stated in the Constitution, there are various acts to support the administrative implementation thereof. The Protection of Personal Information (POPI) Act of 2013 shows similarities to other national legislation such as the Data Protection Act of the United Kingdom, and seeks to protect the right to privacy of 'persons' which in South Africa goes further than in some other countries, because it covers individuals (natural persons) as well as institutions (legal persons). POPI allows for exceptions to strict rules that limit the gathering and retention of personal of data, if such data are used for research purposes. POPI also aims to regulate cross-border transfer of personal information. These requirements are relevant to researchers interested in the analysis of data that contain information on human subjects. POPI is complemented by the Promotion of Access to Information Act of 2000, which supports the constitutional right of access to information (required for the exercise or protection of any rights) that is held by the State or by another person.

South Africa also has a Statistics Act of 1999, which, amongst other things, deals specifically with confidentiality of personal information, which once again includes not only individuals, but also households, organisations or legal entities. This act is currently (2016) being reviewed to specifically deal with the opportunities and challenges of Big Data and the protection of personal information in the national statistical system. An act that explicitly protects the rights of research participants is the National Health Act of 2003. One of the chapters of the act deals with 'health research' and protection for human subjects who participate in such research. The definition of 'health research' is quite broad, and includes research on 'the biological, clinical, psychological or social processes in human beings'. The establishment, roles and functions of health research ethics committees are provided for in the act, and attention is also given to the management of health information, including conditions under which such information may be used for research purposes. Strict conditions are set for research involving vulnerable respondents such as children.

APPENDIX 4. A PRIVACY HEURISTIC (WHAT, WHO, WHERE, WHY)

It is crucial to remember that each of the three conceptual bases of privacy outlined in **Section 4.1** - private/public agency, spheres of privacy, and private information - is unstable and context dependent. The dichotomisation into private and public forms of agency is seriously flawed in that it does not take into account that private corporations are now among the main stakeholders vying for access to private individuals. The notion of spheres of privacy does not help us much in a world where the private sphere is made public by many means, from social media to data brokers. A focus on private information is misguided in that, ultimately, any piece of information might in principle be considered an appropriate object for privacy protection, depending on context or combination.

A more practical approach to a 'privacy reflection tool' might be for the following questions to be posed and answered conscientiously by the researcher. (Some of the issues raised below are relevant not only to researchers, but also to funders and reviewers.) The questions are not meant to indicate that only one reply is the right one, but rather to help identify the salient issues in each case. Since this heuristic reflects the importance of considering the ongoing context/local background and of taking reasonable expectations seriously (see also **Section 4.6**), it can be applied more widely than just to Big Data.

What

In making decisions that concern privacy, it is important to realise that there is no such thing as simply public information. Any bit of information can in principle, given the circumstances, amount to private information. Nevertheless, seriously considering what sort of information one is dealing with is central to a proper evaluation of the legitimacy of dealing with that information. In the New Forms of Data world, what determines data's status as anonymous, personal, or sensitive is the nature of the combined sets of data being used and processed rather than that of the various subsets or sources seen in isolation.

A basic distinction is the mutually exclusive one between *personal* data and *anonymous* data. Personal data is directly tied to an identified or identifiable individual. Anonymous data, by contrast, is any set of data which is non-identifiable, or where the risk of re-identification is deemed minute (requiring an effort not realistically undertaken). The stricter definition of anonymised data means that if a piece of information can be tied to an identified or identifiable individual, then those data are not anonymous.

In some traditions of thinking about privacy (present also in, *e.g.*, European and Australian privacy legislation), a further subset of types of personal information is characterised as *sensitive* personal information. The list typically includes person-identifiable information about race/ethnicity; political/religious/philosophical opinions/beliefs; health; sexuality; and union membership.

Again, the list serves as a reminder that what is considered sensitive information depends on the person, the cultural background, and other political and historical issues. (To some, economic information about salary might be considered more sensitive than information about, *e.g.* union membership – although this is probably not the case if you are a shipyard worker in 1970 Gdansk.) But at the same time, the list does function as a heuristic in trying to consider what the person(s) in question might think of as sensitive information.

Question 1: Is the data to be handled *personal* data? (handling includes collecting the data or passing it on)

Question 2: Is (some of) the data likely to be considered *sensitive* to the persons in question?

Who

Similarly, and sometimes intricately intertwined with the 'What' dimension, the question of whose information is being handled can constitute a highly relevant ethical difference in determining whether – or how – to go through with the activity. One of the legitimising aspects of social science research is that it can serve to reveal aggregated social injustice, unwarranted differences, or abuse of power. All else being equal, people in positions of power should not necessarily expect the same level of discernment-based privacy protection (that is, privacy protection going beyond unambiguously articulated requirements) as those disempowered through, *e.g.* poverty or other forms of relative helplessness, or who are subject to greater threats of discrimination and other forms of unlawful or unfair exploitation.

Question 3: Do the data subjects (data subjects: those whose information is being handled) have the resources to react if they find the handling objectionable?

Question 4: Are the data subjects in positions of power that should make them expect legitimately heightened attention from researchers?

In research, and perhaps especially in aggregate, population-based research, there is a separate responsibility to consider not only individuals but groups. This is at least partly due to the fact that identification or construction of groups in dissemination of results can lead to unwarranted stigmatisation or discrimination of that group, phenomena that amount to unwanted consequences for individuals recognised as belonging to the group – whether thought of in terms of wealth, race/ethnicity, age, gender, nationality, affliction, *etc*.

Question 5: Might the research lead to the unwarranted stigmatisation of, or discrimination against, a group?

The concept of groups can also be relevant to an ethical evaluation of a social science research project involving data subjects that represent groups of people who might rightly be considered particularly vulnerable in some relevant sense. Among the many groups that have been identified as vulnerable are children, immigrants, prison inmates, pregnant women, employees, and ethnic minorities. Again, very few groups are simply vulnerable *per se*: one must look to the context to see whether situations connected with the information or engendered by the research make one or several groups vulnerable in a relevant way. In particular, with New Forms of Data, there is an increased risk of creating new groups for discrimination or of classifying data subjects in ways that perpetuate historical bases for discrimination without the victims being able to contest those classifications[21].

Question 6: Do any of the data groupings constitute or represent, or perhaps even contribute to creating, vulnerable groups of individuals or organisations?

Where

Most of us would consider the socio-physical setting to be highly relevant in determining whether or in what way privacy is to be expected. A few examples of different socio-physical settings are: workplace, home, patient room in hospital, doctor's office, car, train compartment, city street, private bathroom. None of these places are public per se in the specific sense of making public any activity taking place there. (A private conversation taking place on a city street is reasonably expected to remain private nonetheless.) All the same, we have different expectations concerning each of them, constituted partly by shared social notions of which behaviour is fitting in each sphere.

Question 7: What are the places/settings/situations in which the information is gathered, and might the handling be deemed to go against reasonable expectations correlated with those settings?

Why

The evaluation of the 'Where' dimension of an ethical heuristic for the handling of data in social science research depends on the purposes of the research. We understand a setting or a situation partly in terms of its object – its aims or goals - which amount to a concretisation of part of its *values*[22]. A university, say, has its *raison d'être* partly as a place of learning in the sense of being a place where the characteristic activities have as their aims/goals such things as the imparting of learning, the transmission of intellectual tradition, and the upholding of a critical gaze towards issues political, social, and historical.

Question 8: Does the research activity in any way undermine or work against the aims and values of the place/activity/institution from which the data are gathered?

'Why' does not pertain only to the purposes and values of research, however. Research is not outside the world, and research finds a primary legitimisation in acknowledging this fact. Accordingly, the researcher also needs to weigh her own activity in relation to the activities she might be disrupting, transforming, or subverting.

Question 9: How do the purposes and values of the current research project harmonise or conflict with the aims, goals, and values of the sites the research might be affecting?

Question 10: What might be considered the data subjects' reasonable expectations concerning the research project's re-contextualisation of their information?

REFERENCES

Barocas, S. and H. Nissenbaum (2014) 'Big Data's End Run around Anonymity and Consent'. Chapter 2 in J. Lane, V. Stodden, S. Bender and H. Nissenbaum (eds.) (2014) *Privacy, Big Data and the Public Good: Frameworks for Engagement.* Cambridge University Press: Cambridge.

Benhabib, S. (1994). 'Deliberative Rationality and Models of Democratic Legitimacy'. *Constellations* 1 (1): 26–52. doi:10.1111/j.1467-8675.1994.tb00003.x.

Cameron, D. S. Pope, and M. Clemence. (2014). *Dialogue on Data: Exploring the Public's Views on Using Administrative Data for Research Purposes.* Ipsos MORI Social Research Institute. https://www.ipsos-mori.com/DownloadPublication/1652_sri-dialogue-on-data-2014.pdf.

Chambers, S. (1996). *Reasonable Democracy: Jurgen Habermas and the Politics of Discourse*. Ithaca, N.Y.: Cornell University Press.

Council of Europe, European Court of Human Rights (1950) *European Convention on Human Rights*, Rome, November 4.

Crawford, K. and J. Schultz (2014) 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms', 55 B.C.L. Rev. 93, http://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4

ESRC (2015) ESRC *Framework for Research Ethics*. Available at: http://www.esrc.ac.uk/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/

ESSNet SDC (2010) *Handbook on Statistical Disclosure Control*. Available at: http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Farrar, C. J. Fishkin, D. Green, C. List, R. Luskin, and E. Levy Paluck. (2010). 'Disaggregating Deliberation's Effects: An Experiment within a Deliberative Poll'. British Journal of Political Science 40 (2): 333–47.

Global Alliance for Genomics and Health (2014) *Framework for Responsible Sharing of Genomic and Health-Related Data.* Available at: https://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data

Greenwood, D., A Stopczynski, B. Sweatt, T. Hardjono, and A. Pentland (2014): 'The New Deal on Data: A Framework for Institutional Controls', chapter 9 (192-210) in J. Lane, V. Stodden, S. Bender and H. Nissenbaum (eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press.

Habermas, J. (1984). *The Theory of Communicative Action, Volume Two: Lifeworld and System: A Critique of Functionalist Reason.* Boston: Beacon Press.

ICPSR (2015) *ICPSR Glossary*. Available at http://www.icpsr.umich.edu/icpsrweb/datamanagement/support/glossary

Luskin, R., J. Fishkin and R. Jowell. (2002). 'Considered Opinions: Deliberative Polling in Britain'. *British Journal of Political Science* 32 (3): 455–88.

Merton, Robert K. (1973) [1942], 'The Normative Structure of Science', in R. Merton *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago: University of Chicago Press.

Nissenbaum H. (2011) *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford Law Books). Stanford University Press. Stanford, California.

OECD (2013) *The OECD Privacy Framework*. Available at: http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

OECD (2014) *OECD Expert Group for International Collaboration on Microdata Access. Final Report*. Available at: http://www.oecd.org/std/microdata-access-final-report-OECD-2014.pdf

OECD (2015) *Glossary of Statistical Terms*. Available at: http://stats.oecd.org/glossary/

Patil, S., B. Patruni, H. Dunkerley, J. Fox, D. Potoglou, and N. Robinson. (2015). *Public Perception of Security and Privacy: Results of the Comprehensive Analysis of Pact's Pan-European Survey.* RR-704-EC. RAND Corporation. http://www.rand.org/pubs/research_reports/RR704.html .

UK Data Service (2016): *Collections Development Policy*. Available at: https://www.ukdataservice.ac.uk/media/398725/cd227-collectionsdevelopmentpolicy.pdf

UNECE (2009) *Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes* (Geneva). Available at: http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf

UNECE/UNSC (1995) *Guidelines for the Modelling of Statistical Data and Metadata*, Conference of European Statisticians Methodological Material. Available at: http://www.unece.org/fileadmin/DAM/stats/publications/metadatamodeling.pdf

United States Department of Health, Education and Welfare (1978). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. National Commission for the Protection of Human Subjects of Biomedical and Behavioural Research. Washington, DC: United States Government Printing Office.

UUK (2013) *The Concordat to Support Research Integrity*. Available at: http://www.universitiesuk.ac.uk/highereducation/Documents/2012/TheConcordatToSupportResearchIntegrity.pdf

Van Mil, A. and h. Hopkins (2015) *Big Data: Public Views on the use of private sector data for social research*. http://www.esrc.ac.uk/files/public-engagement/public-dialogues/public-dialogues-on-the-re-use-of-private-sector-data-for-social-research-report/

Vomfell, L., F. Stahl, F. Schomm and G. Vossen (2015) 'A Classification Framework for Data Marketplaces'. In: Working Paper No. 23, European Research Center for Information Systems, J. Becker et al. (eds.) Munster 2015. Available at: https://www.ercis.org/sites/ercis/files/structure/network/research/ercis-working-papers/ercis_wp_23.pdf

Walzer, M. (1983) *Spheres of Justice: A Defense Of Pluralism And Equality*, Basic Books.

NOTES

1. http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

2. See Appendix 2 for examples of the types of data that this phrase (as used here) refers to. For clarity, this will be referred to collectively as 'New Forms of Data' throughout the remainder of this report.

3. A key text is Merton (1973).

4. This trifurcation of ethical principles has been explicit in most research ethics since the *Belmont Report (US Department of Health, Education and Welfare 1978).*

5. See, for example, http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/chapter3-chapitre3/

6. See, for example, Joint Research Compliance Office, Imperial College London http://www3.imperial.ac.uk/clinicalresearchgovernanceoffice/researchgovernance/whatisresearchgovernance and several other websites where the same definition is provided.

7. See Section 4.3 for more detail.

8. For instance, the UK-based Economic and Social Research Council (ESRC) and the tri-council panel on research ethics in Canada have publications on research ethics (also touching on new forms of data and privacy issues) available at http://www.esrc.ac.uk/files/funding/guidance-for-applicants/esrc-framework-for-research-ethics-2015/ and http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/

9. *E.g*. a group at the University of Cambridge http://www.crassh.cam.ac.uk/programmes/ethics-of-big-data-research-group and the North America-based Computer Research Association; also see www.cra.org and their white paper at http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf; also a blog managed by a data consulting firm at http://idatassist.com/ethics-policy-inquiry-and-the-positive-impact-on-data-science.

10. *E.g.* Secretary's Advisory Committee on Human Research Protections in the USA, see http://www.hhs.gov/ohrp/sachrp

11. The UK Data Archive refers to five 'safes', see **Section 4.5** for more details.

12. It is difficult to adequately define, or engage with, a grouping as vague as 'the general public' However, public opinion and what Habermas (1984) referred to as 'the public sphere' (*e.g.* society engaged in critical public debate) remain important role players in any governance system, including the governance of research. There are also other 'publics' whose concerns need to be accounted for.

13. See *The UK Administrative Data Research Network: Improving Access for Research and Policy. Report from the Administrative Data Taskforce*. Swindon. Economic and Social Research Council, p. 2, 2012. http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/

14. For further discussion, see UK House of Commons Science and Technology Committee, *Responsible Use of Data*, 19 November 2014.

15. See http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (2012/0011 (COD)).

16. For example, see http://sagebase.org/e-consent/

17. An example relates to the release of all of New York's yellow cab journeys data for 2013 in response to a Freedom of Information request. While the identity of each cab was anonymised, it was possible to locate similar journeys between two points, matching these with Spokeo and Facebook location data to reveal the identities of specific passengers who had travelled from known origins. See http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/ for details.

18. Two examples are *differential privacy* and *k-anonymity*.

19. http://www.olswang.com/eu-data-protection-reform/privacy-by-design/

20. The trust envisaged is thus not mere blind faith, but a more reflective and experience-based trust analogous to that expounded (for a much more formalised and limited forum) in OECD (2014).

21. Crawford and Schultz (2014).

22. Cf. Walzer (1983).

ABSTRACT

This report sets out some basic rules that underpin an ethical approach to research using new forms of data for social and economic research. These rules and the interpretation that we place upon them give rise to a set of recommendations designed to provide a framework for the ethical governance of research using such data. There are assumptions and limitations underpinning these recommendations – they are not cost-free and will be easier to apply in countries with established research ethics procedures, particularly where research organisations and data owners have access to ethical review bodies. The sharing of expertise on, and knowledge about, research ethics between countries is critical to the creation of a common and cost-efficient ethical environment for social scientific research.

Keywords: Research, ethics, data.