Science & Technology Committee Inquiry - The right to privacy: Digital data

Response from the British Academy

Contents

About the British Academy

The British Academy is the UK's national academy for the humanities and social sciences. We mobilise these disciplines to understand the world and shape a brighter future. From artificial intelligence to climate change, from building prosperity to improving well-being – today's complex challenges can only be resolved by deepening our insight into people, cultures and societies. We invest in researchers and projects across the UK and overseas, engage the public with fresh thinking and debates, and bring together scholars, government, business and civil society to influence policy for the benefit of everyone. The Academy, alongside the other national academies, is a distinctive element in the research funding ecosystem and complementary to UKRI, offering responsive, bottom-up grants at key career stages, from early career to senior fellowships.

Introduction

In this response, the British Academy will focus on responding to the below topic within the Science & Technology Committee's Inquiry:

- the potential benefits, including to research, to effectively use and share data between and across Government, other public bodies, research institutions and commercial organisations, and the existing barriers to such data sharing.

The response consists of an executive summary of key insights drawn from the materials in this response, and from previous British Academy outputs on this topic, such as reports and roundtables.[1] Following the executive summary are two viewpoints on the topic of researcher access to data, authored by Fellows of the British Academy Professor Anna Vignoles and Professor David Hand. The final section features a summary of some of the insights that emerged from a roundtable on data reform that the British Academy held at the end of 2021 in collaboration with the other National Academies (in response to the government's *Data: a new direction* consultation).[2]

Executive Summary

The opportunity to link data represents a vital opportunity to undertake research in SHAPE disciplines (Social Sciences, Humanities and the Arts for People and the Economy) that can lead to better lives for citizens. Linking anonymous administrative datasets from across government, industry, and surveys, holds tremendous potential for research leading to individual, societal and economic gains.

Linked data enable researchers to investigate questions that cannot be addressed by individual data sets, across a range of areas in society such as health, wellbeing, education, and crime. However, it is essential that the right to privacy is adequately addressed in doing so, so that research is undertaken in a demonstrably safe and transparent way that maintains public trust, whilst making the most of opportunities to connect data sets for social good.

Technical and **social** strategies to protecting privacy must be thought of as going hand in hand:

- **Technical strategies** include the use of a range of technologies in the context of research for minimising privacy breaches: including pseudonymisation and anonymisation, secure multiparty computation, homomorphic computation, and blockchain technologies. They also include maintaining good practice such as proper data storage and data analysis being conducted only within secure research environments. The ONS 'five safes' approach illustrates what can be done in this regard (Stokes 2017). More recently, the ONS Integrated Data Service is a welcome example of a valuable way to build upon the work of the Secure Research Service (ONS 2021a).
- Equally critical are **social strategies** to ensure the public are made aware of the potential benefits of sharing data. Studies have shown that the public are much more accepting of the risks to privacy when the benefits of sharing data are made clear, especially the value of data sharing for the individual.

[1] These viewpoints represent the views of the individuals named and are not necessarily formal positions of the British Academy.
[2] The British Academy and the Royal Society (2017) have previously examined questions of data governance more extensively in the joint report *Data management and use: Governance in the 21st century*.

- These should be coupled with strategies to **shift culture** among data owners away from a risk averse and overly cautious culture toward one that emphasises the social purpose and benefit of data for the public.
- **Legislation** needs to be explicit that using data for research purposes is both permitted under law and facilitated where possible. Some aspects of current legislation, such as the *data minimisation strategy* in the GDPR, can potentially undercut the promise for society arising from data mining and linkage.
- Finally, more **investment** in long-term data infrastructure will enable the UK to benefit significantly from its data linkage. Crucially, this means investment in *processes, people* and *institutions*, as well as technical *infrastructure* such as storage, curation and cleaning.

Finally, alongside opportunities relating to the use of government data, there are increasing volumes of data held by the **private sector** – in particular, technology companies:

- It will be valuable for research and indeed for the state to have insight that can be gained from anonymised use of this data. For instance, we have already seen how Transport for London has used mobile phone Wi-Fi data to understand commuter flows (BBC 2019).
- Large technology companies have often not been that open to letting researchers access data for research purposes. It will be helpful for the inquiry to consider ways in which this kind of data could be accessed for public good.
- One option might be to enable people to ‘donate’ anonymised data about themselves for research – technology companies could make this option available (Orben 2021). Another option might include the creation of a public interest trust which holds such data as a ‘data commons’ after a period (Shah 2018).
- The Digital Economy Act has enabled the ONS to have better access to private sector data for statistical purposes – it would be useful for the committee to review how this has worked since the Act was put into place.

Data use and privacy

Professor Anna Vignoles CBE FBA

The potential benefits for UK research from using linked anonymous data (from government, industry and surveys) cannot be understated. Linked administrative data have enabled the New Zealand government to understand who the (small) group of heavy users of government services are and how policies might reduce this welfare need (ADR UK 2021). In the UK, linked education data have enabled us to understand better the labour market outcomes from particular educational routes (Anderson & Nelson 2021). More extensive linkage could, for example, help us quantify the long run return on investing in children in their early years, in terms of lower welfare costs in adulthood and reduced burden on the criminal justice system. The potential individual, societal and economic gains from using such data for research are huge.

Legislation needs to be explicit that using data for research purposes is both permitted under law and facilitated where possible. However, we must maintain public confidence that data is being used for social good.

To fully realise the benefits of data and data linkage, there are three factors to consider. First, privacy must be protected, and public trust must be built. Second, the culture among data owners needs to change. Third, more investment in data infrastructure (technical and human) is needed.

1) **Privacy**. While the risk of data breaches cannot be eliminated, we have the technical tools to minimize this risk. The ONS ‘five safes’ approach (Stokes 2017) illustrates what can already be done. Novel technologies will no doubt be useful in the future but risk minimization is already possible. A bigger problem is public confidence and trust. We need to increase the public’s appreciation of the benefits of using data for research and their understanding of how anonymous data can be used safely. This applies to all types of data, including health.

   To build trust, the distinction between anonymous data and data that can identify individuals must be very clear.

   There is a distinction between research for the social good (which may have commercial/economic value) and that for immediate commercial gain. However, the line between these is often blurred. We must have robust processes and institutions capable of distinguishing between different uses of data, applying the necessary controls and ensuring that the ethical framework reflects the views of the public.

2) **Culture**. There has been resistance in some parts of government to linking data across departments, even where legal barriers do not exist. This overly cautious culture partly reflects a reluctance to share information but is also driven by the fact that data holders are personally responsible for data breaches, which makes them individually risk averse. Subject to the issues raised in (1), the default assumption needs to be that data held by public bodies will be shared unless there is a demonstrably good reason not to.

3) **Investment**. Data is collected at great expense and is a major asset. Insufficient investment in preserving the long-run value of data undermines its usefulness. This

requires investment in data cleaning, curation, access and skills, with a presumption that preserving data for long term use by a range of researchers is the most ethical approach, given the time that individuals and institutions have put into creating the data in the first place.

Researcher access to government data

Professor David J. Hand OBE FBA

When data mining was created as a formal discipline, some 20-30 years ago, it was premised on the truth that there was huge value latent in existing large data sets. With the continued growth in numbers and size of data sets, further opportunities have arisen over the past twenty years. In particular, the *linkage* of data sets from diverse sources has huge promise for benefiting society when coupled with data mining. Linked data permit us to answer questions which could not be addressed by individual data sets, ranging over all aspects of human society, including health, education, crime, and every other aspect.

However, concomitant on this huge promise is a risk that the individuals can be identified and their privacy compromised. This is a real risk, since one cannot predict with certainty what data sets might be created and merged in the future. This means that anonymisation cannot be guaranteed: there is always some possibility of individuals being identified. This is a manifestation of the more general truth that risk cannot be reduced to zero. However, in the context of research, various technologies for minimising risk have been developed. These include pseudonymisation and anonymisation, secure multiparty computation, homomorphic computation, and blockchain technologies. They also include data analysis being conducted only within secure research environments, with the data not allowed to leave the premises or storage facility.

Possibly even more important than those technical strategies, however, it is critical that the public are made aware of the potential benefits of sharing data. Studies have shown that the public are much more accepting of the risks to privacy when the benefits of sharing data are made clear. This is perhaps illustrated by the "public/private data donation paradox", in which people readily divulge information to commercial bodies, but are more hesitant about doing so to official bodies. The explanation lies in the fact that the benefit from the data exchange in the commercial case is immediate and apparent; namely that a service or good is received. One aspect of this is that the value of the data for the individual must be made clear. It is not enough to say "society will benefit", without spelling out how individuals will benefit.

Another strategy aimed at tackling the risk has been widely adopted. This is the *data minimisation* strategy contained within the GDPR. GDPR Article 5(1)(c) defines data minimisation as the collection of data that are "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed”. Experian defines data minimisation as "a principle that states that data collected and processed should not be held or further used unless this is essential for reasons that were clearly stated in advance to support data privacy." As will be apparent, this strategy works to undercut the potential promise for society arising from data mining ("the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners", Hand, 1998a,b) and from data linkage ("Data linkage has never been more important for our society", ONS, 2021b).

Note on Insights from National Academies Roundtable on Data Reform

In response to the government consultation on data reform, *Data – A New Direction*, the National Academies held a roundtable to engage experts from the four academies and government representatives in a discussion of the government's proposals on GDPR reform with regards to data for research and innovation. Discussions were centred around impacts GDPR has had on researcher access to data, and how successfully it has been implemented and interpreted. Further topics included other barriers to data access for research, and the variety of technologies and tools that could be used to support researcher access while mitigating other concerns such as data security and privacy. The Royal Society's submission to the *Data – A New Direction* consultation (2021) covers some of the points raised at this roundtable in greater detail.

Key Points

GDPR has served as an impediment to researchers accessing data, repurposing it, and linking data sets. This is primarily due to the culture of interpretation that surrounds GDPR, rather than the text itself. The regulations already make provision for greater researcher access, but use of these is limited by a lack of precision in terminology, a culture of deny-access-by-default and a lack of clear guidance.

- Reasons for this risk-averse culture include:
  - A lack of clear guidance on researcher access
  - The personal liability of data holders in the case of data breaches
  - GDPR being used as an excuse when there are other motivations for denying researcher access.
  - In some cases, researchers (e.g. in humanities/social sciences) have to grapple with guidance designed for different disciplines to their own (e.g. medical sciences)

- **Public trust and confidence** in the regulatory system are key to greater researcher access and innovation.
  - Language that implies any regulatory reform is done primarily with the aim of loosening protections is liable to undermine this.
  - Lack of enforcement activity on the ground also limits public trust. In the last ICO annual report there were only three fines and one Enforcement Notice, despite over 30,000 complaints, with no detail as to why.

- The system should enable **individuals to allow the transfer of privately-held data** about them to researchers in the public sector. This could facilitate access to high quality data for researchers.

- The present data access regime benefits large companies with the ability to collect large amounts of personal data, while disadvantaging public research done for public benefit, and SMEs that could help drive responsible innovation. Government has a role in brokering data sharing between such private and public actors.

- Trusted Research Environments and pseudoanonymisation of data are effective means of protecting privacy in shared data and **limiting the risk of data breaches**. The development of Privacy Enhancing Technologies (PETs) will improve this even further. However, this risk can never be totally eliminated, and focus on it should be balanced against potential benefits.

- Sharing data with researchers in Europe is essential and maintaining **data adequacy** with the EU is therefore important.
    - There may be scope for some regulatory divergence while maintaining data adequacy. The EU's existing data adequacy agreements suggest this, though these were grandfathered into the system and so may not be a sound guide.

References

ADR UK (2021). New Zealand's Integrated Data Infrastructure: Linking data for better science and policy.

Anderson, Oliver, and Nelson, Moira, Department for Education (May 2021). Post 16 education and labour market activities, pathways and outcomes (LEO). Research report. Government Social Research.

BBC (2019). TfL to track tube users via their wi-fi to ease overcrowding.

The British Academy and the Royal Society (2017). *Data management and use: governance in the 21st century*.

Hand D.J. (1998a) Data mining - reaching beyond statistics. *Research in Official Statistics*, **2**, 5-17.

Hand D.J. (1998b) Data mining: statistics and more? *The American Statistician*, **52**, 112-118.

ONS (2021a) ONS launches Integrated Data Service to boost government collaboration on data sharing.

ONS (2021b) Joined up data in government: the future of data linking methods.

Orben, Amy (2021). Written evidence submitted by Dr Amy Orben College Research Fellow at the University of Cambridge (OSB0131).

Royal Society (2021). Royal Society Submission to the Data: A New Direction consultation.

Shah, Hetan (2018). Use our personal data for the common good. Nature, World View.

Stokes, Peter (2017). The 'Five Safes' – Data Privacy at ONS. ONS blog.