Office for Statistics Regulation

Data Sharing and Linkage for the Public Good

July 2023

Executive Summary

Data sharing and linkage in government stands at a crossroads.

Since the Office for Statistics Regulation (OSR) last reported on this, there has been some excellent progress in creating linked datasets and making them available for research, analysis and statistics. The pandemic provided a particularly strong impetus to share data for the public good. But, despite the value of sharing and linking data being widely recognised, there remain areas of challenge and uncertainties about the public's attitude to, and confidence in, data sharing. Unless significant changes are implemented, we are concerned the progress that has been made could be lost and the potential for data sharing and linkage to deliver public good will not be achieved.

Our report focuses on data sharing and linkage within central government and the devolved administrations of the UK[1], but most of our findings could also be relevant for the wider public sector. We plan to use the report as a platform to engage across and beyond government, working with others to help realise our recommendations and to help government enhance the public good of data and statistics.

Our review

Our review takes stock of data sharing and linkage across government. It points the way to build on recent successes and to confront the more ingrained challenges. To help others within government see how barriers can be overcome, and to enable positive action, we draw on inspirational examples of data sharing and linkage. Our findings and recommendations relate to four overarching themes:

1. **Public engagement and social licence[2]:** The importance of obtaining a social licence for data sharing and linkage and how public engagement can help build understanding of whether/how much social licence exists and how it could be strengthened. We also explore the role data security plays here.

2. **People:** The risk appetite and leadership of key decision makers, and the skills and availability of staff.

3. **Processes:** The non-technical processes that govern how data sharing and linkage happen across government.

4. **Technical challenges:** The technical specifics of datasets, as well as the infrastructure to support data sharing and linkage.

[1] During our report we use the term 'government' to refer to the UK government and the devolved administrations of Wales, Scotland and Northern Ireland. Where we wish to refer to a specific administration, this is made clear.

[2] Social licence, in this context, refers to the level of acceptance or approval by the local communities of interest and/or stakeholders for a data sharing and/or linkage project.

Learning from future scenarios for data sharing and linkage

To help individuals and organisations within government explore the possible implications of the choices they make now about whether to share or link data, we consider four possible 'future scenarios' for data sharing and linkage, set five years from now. They are not predictions but stylised versions of possible futures, which help to bring out the impact on public good of acting on (or not acting on) the current barriers that exist to data sharing and linkage. The scenarios are:

1. **Data sharing and linkage for the public good:** This is the best-case scenario we currently see. Under this scenario, data sharing and linkage is a priority across government and many of the barriers identified in our report have been removed. This is achieved through high levels of collaboration among organisations within government, as well as strong partnerships with external researchers and organisations. In this scenario, opportunities to enhance the public good of data and statistics are fully realised and missed data use is very rare.

2. **Data sharing and linkage in silos:** In this scenario public good is being realised in certain topic areas, but opportunities are being missed in other areas. In the pockets where things are going well, senior leaders are proactive and engaged, collaboration is high and consistency of practices helps things run smoothly. However, there are pockets where little to no progress is being made.

3. **Data sharing and linkage for government:** This scenario features high levels of cooperation across government organisations but low engagement with researchers and organisations beyond government. As a result, the value of providing access to data to external researchers, and of sharing outputs of analysis beyond government is not being considered or realised. In this scenario, public good is not being realised and there is considerable missed use of data.

4. **Data sharing and linkage deprioritised:** This is the worst-case scenario we envisage, where there has been a breakdown in support for data sharing and linkage and progress previously made across government has been lost. There are many examples of missed opportunities where data could have a real impact and, consequently, the potential for data sharing and linkage for the public good has not been realised.

The contrasts between these scenarios bring into greater focus the costs associated with not sharing and linking data, as well as the benefits when it is done. Based on this exploratory thinking and interview findings, we make **16 recommendations** that, if realised, would move government away from the three less desirable scenarios and towards enabling greater data sharing and linkage for the public good. To help generate and maintain momentum on these recommendations, OSR will review and publicly report on progress towards them between six months and one year after this publication.

Our recommendations

To enable greater data sharing and linkage for the public good through tackling current barriers we make the following recommendations. High-level findings are also provided for context.

Public engagement and social licence

Key findings

- There is a need for more public engagement about data sharing and linkage, to improve both transparency of work that is being carried out, and public confidence in data sharing and linkage more generally.
- There is growing evidence that some people in the UK want and expect their data to be used when it is done securely and transparently.
- While there are examples of public engagement is being done well, there can also be a lack of understanding about how to do public engagement effectively.
- The Public Engagement in Data Research Initiative (PEDRI) is a new sector-wide partnership looking to bring together organisations who work with data and statistics to collaborate and embed meaningful public involvement across the data ecosystem. This initiative could strengthen the public engagement landscape, sitting alongside other existing centres/initiatives that already support specific communities.
- The amount of social licence for a data sharing or linkage project is likely to be related in part to data security. The Five Safes Framework is a set of principles employed by many data services that enable them to provide safe research access to data. Assurance that the Five Safes Framework continues to support the appropriate level of security would be helpful.
- Privacy Enhancing Technologies (PETs) are newer technologies that can help organisations share and use people's data responsibly, lawfully and securely. There is growing interest in PETs and the benefits their use could bring.

Recommendations

Recommendation 1: Social Licence: The government needs to be aware of the public's views on data sharing and linkage, and to understand existing or emerging concerns. Public surveys such as the 'Public attitudes to data and AI: Tracker survey' by the Centre for Data, Ethics and Innovation (CDEI) provide valuable insight. They should be maintained and enhanced, for example to include data linking.

Recommendation 2: Guidelines and Support: When teams or organisations are undertaking data sharing and linkage projects, there is a growing practice of engaging with members of the public to help identify concerns, risks and benefits. To help teams or organisations who are undertaking public engagement work, best practice guidelines should be produced, and support made available to help plan and coordinate work. This should be produced collaboratively by organisations with experience of this work for different types of data and use cases and brought together under one partnership for ease of use. We consider that, given its current aims, the Public Engagement in Data Research Initiative (PEDRI) could be well placed to play this role.

Recommendation 3: The Five Safes Framework: Since the Five Safes Framework was developed twenty years ago, new technologies to share and link data have been introduced and data linkage of increased complexity is occurring. As the Five Safes Framework is so widely used across data access platforms, we recommend that the UK Statistics Authority review the framework to consider whether there are any elements or supporting material that could be usefully updated.

Recommendation 4: Privacy Enhancing Technologies: To enable wider sharing of data in a secure way, government should continue to explore the potential for Privacy Enhancing Technologies (PETs) to be used to enhance security and protect privacy where data are personally identifiable. The Office for National Statistics (ONS) Data Science Campus is well placed to lead and coordinate this work.

People

Key findings

- At every step of the pathway to share and link data, the people involved, and their skills and expertise, are instrumental to determining whether projects succeed or fail.
- A big barrier to data sharing and linkage for some organisations is whether it is a priority for the Accounting Officer. Making secure data sharing and linkage a strategic priority at the level of the Accounting Officer in more organisations would enable more joined up approaches across government. To achieve this, an appreciation of the potential benefits needs to be more widely held.
- Recruiting and retaining people with the skills needed to link, maintain and analyse data was a significant challenge raised by many of our interviewees.

Recommendations

Recommendation 5: Data Literacy in Government: To gain the skills to create and support a data-aware culture, it is important for senior leaders to have awareness of and exposure to data issues. One way to raise awareness and exposure would be for senior leaders to ensure that they participate in the Data Masterclass delivered by the ONS Data Science Campus in partnership with the 10 Downing Street (No10) Data Science Team.

Recommendation 6: Data Masterclass Content: The Data Masterclass could expand its topics to include sections specifically on awareness of data linkage methodologies, the benefits of data sharing and linkage and awareness of different forms of data. This would fit well under the Masterclass topics of ‘Communicating compelling narratives through data’ or ‘Data-driven decision-making and policy-making’.

Recommendation 7: Arbitration Process: To facilitate greater data sharing among organisations within government, a clear arbitration process, potentially involving ministers, should be developed for situations in which organisations cannot agree on whether data shares can or should occur. Developing such an arbitration process could be taken on by the Cabinet Office, commissioned by the Cabinet Secretary and delivered working with partners such as No10 and ONS.

Recommendation 8: Career Frameworks: To enable more effective and visible support for the careers of people who work on data sharing and linkage, those responsible for existing career frameworks under which these roles can sit, such as the Digital Data and Technology (DDaT) career framework and the Analytical Career Framework, should ensure skills that relate to data and data linkage are consistently reflected. They should also stay engaged with analysts and professionals across government to ensure the frameworks are fit for purpose. These frameworks should be used when advertising for data and analytical roles and adopted consistently so that career progression is clear.

Processes

Key findings

- There is variation within government over how much data holders and researchers understand the process to share data under different legal bases.
- When applying for data through a secure data platform, the process is often lengthy and can appear overly burdensome to researchers.
- For every data share there will be many teams involved, within the same organisation or from many different ones. Not getting these teams together at the very start can cause major delays to data sharing.
- When researchers have a question about a dataset or process, it can be a challenge to find the right person who can help.
- Funding structures across government tend to be set up so that each department controls its own spend, making successful funding highly dependent on the priorities and vision within each department. This siloed approach is hampering efforts of collaboration and means projects with external funders are often more successful.

Recommendations

Recommendation 9: Overview of Legislation: To help researchers understand the legislation relevant to data sharing and linkage and when it is appropriate to use each one, a single organisation in each nation should produce an overview of legislation that relates to data sharing, access and linkage, which explains when different pieces of legislation are relevant and where to find more information. This organisation does not need to be expert in all legislation but to be able to point people to those that are. The Office for Statistics Regulation (OSR) will help convene those in this space to understand more about who might be best placed to take this on.

Recommendation 10: Broader use cases for data: To support re-use of data where appropriate, those creating data sharing agreements should consider whether restricting data access to a specific use case is essential or whether researchers could be allowed to explore other beneficial use cases, aiming to broaden the use case were possible.

Recommendation 11: Communication: To ensure data application processes are fit-for purpose and well understood, those overseeing accreditation and access to data held in secure environments should prioritise ongoing communication with users, data owners and the public to explain and refine the information required. Wherever

possible, they should offer face-to-face or virtual discussions with those applying to access data early in the process, to ensure clarity around both the data required and the process to access it.

Recommendation 12: Checklists: To ensure all necessary teams are involved at the outset of a data sharing and linking project, organisations should consider the use of a checklist for those initiating data sharing. The checklist should contain all contacts and teams within their organisation who need to be consulted to avoid last minute delays.

Recommendation 13: Transparency: Every organisation within government should be transparent about how the data they hold can be accessed and the process to follow. This guidance should be presented clearly and be available in the public domain with a support inbox or service for questions relating to the process.

Recommendation 14: Funding Structure: To allow every organisation a consistent funding stream for their projects, a centralised government funding structure for data collaboration projects across government, such as the Shared Outcome Fund, should be maintained and expanded.

Technical challenges

Key findings

- It can be a challenge for those linking data to get enough information about the data to provide a high-quality linked output with a measurable rate of error.
- While we heard many positive reflections on the effectiveness of current data linkage methodologies, and the way that these are being developed, it was also acknowledged that methodological challenges do still exist, which can also themselves lead to issues with the quality of linked data.
- Variation in data standards and definitions used across government is making linking harder.

Recommendations

Recommendation 15: Sufficient resources: To enable effective, efficient, and good quality data linking across government, senior leaders should ensure there are sufficient resources allocated to developing quality metadata and documentation for data held within their organisations.

Recommendation 16: Standardisation: Many departments are looking to standardise government data and definitions, but it is unclear whether or how these initiatives are working together. Those working to standardise the adoption of consistent data standards across government should come together to agree, in as much as is possible for the data in question, one approach to standardisation which is clear and transparent. Given the work done by the Data Standards Authority, led by the Central Digital and Data Office (CDDO), the CDDO may be best placed to bring this work together.

Office for Statistics Regulation (OSR)

We provide independent regulation of all official statistics produced in the UK. Statistics are an essential public asset. We aim to enhance public confidence in the trustworthiness, quality and value of statistics produced by government.

We do this by setting the standards they must meet in the Code of Practice for Statistics. We ensure that producers of government statistics uphold these standards by conducting assessments against the Code. Those which meet the standards are given National Statistics status, indicating that they meet the highest standards of trustworthiness, quality and value. We also report publicly on system-wide issues and on the way statistics are being used, celebrating when the standards are upheld and challenging publicly when they are not.

Acknowledgements

We would like to acknowledge the significant input of all the individuals and organisations that contributed to this review as we gathered information and tested the ideas presented here. Organisations and teams that contributed are listed in Annex B.

Contents

Introduction

Data sharing and linkage for the public good

Every day, government organisations generate data that have the potential to serve the public good. These data can hold the key to understanding and answering society's most pressing questions. Within government, data can inform the delivery of vital public services, policy developments, evaluation, and answer valuable research questions. Beyond government, data can be a powerful tool that enables organisations and individuals to hold the government to account and to make their own decisions.

When data are shared and linked across government this potential is magnified, enriching insights into society, stimulating innovation and ultimately enabling data, and government, to better serve the public good. Opening up access to data beyond government can significantly increase the analytical capacity to use them for public good, whether it is by feeding evidence back into government or through allowing organisations to make their own decisions.

There are powerful examples that illustrate the value of sharing and linking data across multiple sectors. For example, the Office for National Statistics (ONS) recently published statistics on sociodemographic inequalities in suicides. ONS linked demographic and socioeconomic data about individuals from the 2011 Census with death registration data and, for the first time, was able to show estimates for rates of suicide across a wide range of different demographic groups. ONS believes this analysis will support the development of more effective suicide prevention strategies. Further examples come from Data First, an ambitious data-linking programme led by the Ministry of Justice (MoJ) and funded by Administrative Data Research UK (ADR UK) (see Box 1). Data First aims to unlock the potential of MoJ data by linking administrative datasets from across the justice system and enabling accredited researchers, from within government and academia, to access the data. Data First is also enhancing the linking of justice data with data from other government departments, such as the Department for Education, where linking data has unlocked a wealth of information for researchers about young people who interact with the criminal justice system. Also in the education space, ADR Northern Ireland has recently launched the Education Outcomes Linkage (EOL) 2018/19, a longitudinal database comprised of post-primary schools' data in Northern Ireland, delivered in partnership with the Department of Education and the Department for the Economy in Northern Ireland. A key feature of this project has been prioritising stakeholder and researcher engagement from design to completion, helping to ensure EOL can maximise its goal to drive policy focused research. Finally, the pandemic also provided examples of data sharing and linkage that improved public understanding of the differential impacts of COVID-19 on various population groups.

All these initiatives demonstrate how data sharing and linkage can deliver insights that enable the design of policies that better serve vulnerable groups of society. By looking at them from a different angle and considering the loss if they had not been possible, they also serve to illustrate the enormous cost of missed opportunity if data are not shared or linked, especially when preparing to respond to a national crisis.

What do we mean by data sharing and linkage?

The concept of data sharing is relatively straightforward and involves data normally created in one business area or organisation moving to another. Within this report, we also talk about data access, which reflects the fact that a lot of data sharing across government is now achieved via organisations contributing data to IT platforms, such as databases or modern cloud repositories. These platforms can then enable access to multiple others, removing the need for the source organisation to repeatedly share data themselves. Data sharing and data access often rely on organisations having a common purpose and arrangements, such as an agreement to share data.

Data linkage is the process of joining datasets together. Data that are shared between organisations are often shared with the intention of linking them to further datasets to enhance or improve the data. Data sharing and data linkage are often considered together in this report but, where distinctions exist, these will be made clear. Both data sharing and data linkage come with challenges, which this report will explore.

Why is OSR reporting on data sharing and linkage now?

At the Office for Statistics Regulation (OSR) we see the immense value of data sharing and linkage for decision makers and the wider public. As the independent regulator of the UK's statistical system, OSR is an advocate and a champion for data sharing and linkage, when this is done in a secure way that maintains public trust. It is our ambition that sharing and linking datasets, and using them for research and evaluation, will become the norm across the UK statistical system.

OSR has a unique perspective in sharing and linking data: our vision that statistics should serve the public good means that we have a focus on the availability of data, and analysis that draws on linked data, for individuals and organisations working outside of the government, which others may not. We see a role for OSR as champions of wider value, ensuring opportunities and benefits from linked data are realised by groups beyond government and the public sector, such as academic researchers, so that they can better serve the public good.

In 2018 we published our report Joining Up Data. We identified six key outcomes necessary to achieve a safe and effective data linkage system, underpinned by Trustworthiness, Quality and Value, the three pillars of our Code of Practice for Statistics. In 2019, we published an update report in which we were able to identify progress towards achieving those six key outcomes in several areas.

Since then, there have been several notable changes within the data sharing landscape that have helped to accelerate developments within the statistical system. For example, the powers given to the statistical system via the Digital Economy Act 2017 are now more embedded and have helped to unlock and facilitate access for data sharing and linkage. In addition to this, the Integrated Data Service (IDS) (see Box 2) is being developed as a cross government and researcher service, allowing coordinated and secure access to data for the public good. Meanwhile, strong collaboration between the UK statistical system and ADR UK has supported linkage and sharing of administrative datasets within and across organisations in all four UK

Nations and is helping to make them available to accredited researchers within and beyond government in a safe and secure way.

The coronavirus (COVID-19) pandemic has also had a huge impact on the data sharing and linkage landscape. In response to the desire to answer societal questions concerning COVID-19, the statistical system showed an agile, collaborative and willing approach to sharing and linking data on many topics, including aspects of health, crime, income and housing, spreading across both the private and public sectors. During this unprecedented period, the willingness to share data was driven by a common purpose – the desire to help vulnerable people and ultimately to save lives.

Finally, there is growing evidence that people in the UK want, and expect, data to be used when it is done securely and transparently. There is an expectation by some among the public that their data are already being shared and linked within the public sector for the public good. This position is explored by Data and Analytics Research Environments UK (DARE UK) in its blog 'Trustworthiness of sensitive data research is about more than just privacy and security'.

While there has been some excellent progress in creating linked datasets and making them available for research, analysis and statistics, data sharing and linkage within the government sector now stands at a crossroads. Despite progress, we know there remain areas of challenge around sharing and linkage, and around wider access to data to researchers outside government; and there is a lack of awareness of and uncertainties about the public's attitude to and confidence in data sharing and linkage. These and other areas of challenge have been highlighted by organisations both within and beyond government, including the Social Mobility Commission and the Institute for Government respectively.

This report

This report focuses on how we can empower government to prioritise data sharing and linkage for research purposes, enabling greater data sharing and linkage for the public good. It takes stock of the data sharing and linkage being done for research across government and points the way to build on recent successes and confront the more ingrained challenges.

To inform our position we spoke to stakeholders from across the public sector, including government departments, cross-government linkage projects, trusted research environments (TREs), devolved administrations, data partnerships and government researchers. We explored current barriers to data sharing and linkage from their perspectives and sought to understand opportunities and hopes for the future. Full discussion of our methodology is given in Annex A.

In Chapter 1 we discuss the findings of our interviews with stakeholders, which were conducted between September 2022 and January 2023. This includes the barriers and opportunities that exist in this area, examples of success stories and what can be learnt from them, and further areas of interest that we deem important to understanding the landscape of data sharing and linkage. We make 16 recommendations that, if realised, would enable greater data sharing and linkage for the public good.

Chapter 2 looks to the future of data sharing and linkage in government and presents four possible 'future scenarios' for data sharing and linkage, set five years from now. We illustrate how our recommendations from Chapter 1 can indicate a pathway to data sharing and linkage for the public good.

Information Box 1: Administrative Data Research UK (ADR UK)

ADR UK is a UK data partnership, funded by UK Research and Innovation (UKRI), with a mission to transform the way researchers access the UK's wealth of public sector data, to enable better informed policy decisions that improve people's lives.

The partnership is coordinated by a UK-wide Strategic Hub, which manages a dedicated fund for commissioning research using newly linked administrative data. The Strategic Hub, along with the other ADR partnerships, also promotes the benefits of administrative data research to the public and the wider research community and engages with governments of the UK to secure access to data.

Information Box 2: The Integrated Data Service (IDS)

The IDS is a cross–government project, led by the Office for National Statistics (ONS). It builds on the ONS Secure Research Service, which has been providing secure access to de-identified, unpublished data to accredited researchers for over 15 years. The IDS is a central platform that provides access to data, analytical and visual tools in a secure multi-cloud infrastructure. It aims to be the single data analysis and dissemination platform within government by providing secure and co-ordinated access to a range of high-quality data for government analysts, devolved administrations and external accredited researchers.

In March 2023, IDS entered its Public Beta phase, which marks a step forward in achieving the vision of bringing together ready-to-use data for the public good by expanding the IDS user base and functionality, and offering additional data for analysis, on a safe and secure platform. Data assets available through the platform are listed on the IDS website.

Chapter 1: The current data sharing and linkage landscape across government

An emerging theme from our stakeholders was the overall willingness to share and link data across government and public bodies. The benefits and value of doing this are widely recognised. There is, however, still a wariness around the legality and ethics of data sharing and linking as well as many different processes and ideas around how it should be achieved, which are causing delays.

The picture is not the same in every area of the government. Some areas have moved faster than others and we have found that culture and people are key determinants of progress. Throughout this report, we highlight examples of data sharing and linkage that demonstrate a positive impact for the public good. Through sharing these examples, we hope to enable others to see how barriers can be overcome to take positive action.

To bring out the findings from our interviews, we focus on 'themes' and how these themes were spoken about in the context of both barriers and opportunities. This helps bring out different views and opinions and means there is more opportunity for capturing the complex nature of the landscape. The themes we identified focus on:

- Public engagement and social licence: The importance of obtaining a social licence for data sharing and linkage and how public engagement can help build understanding of whether/how much social licence exists and how it could be strengthened. We also explore the role data security plays here.
- People: The risk appetite and leadership of key decision makers, and the skills and availability of staff.
- Processes: The non-technical processes that govern how data sharing and linkage happens across government.
- Technical: The technical specifics of datasets, as well as the infrastructure to support data sharing and linkage.

Public engagement and social licence

One of the biggest topics mentioned throughout our interviews was the need for more public engagement about data sharing and linkage. Many of our interviewees made a connection between lack of understanding of public perception and the nervousness that still exists around data sharing among some senior leaders, who are concerned by the potential for public resistance. The Office for Statistics Regulation (OSR) agrees with the need for further public engagement, seeing that public engagement serves two important purposes: firstly, to understand and, secondly, to potentially increase the amount of social licence a data sharing or linkage project has.

We have also found through our interviews that the data owners link their perception of social licence for a data sharing or linkage project to data security. In this section, both public engagement and data security will be discussed in the context of social licence. While we sometimes refer to 'the public', we acknowledge that there are different groups that sit underneath this, and it is often useful (and indeed necessary) to engage

with specific groups depending on the topic area and intended outcomes of individual pieces of research.

Public engagement

Our interviews revealed a consensus that those working on data sharing and linkage need to prioritise public engagement in their work, to improve both transparency of work that is being carried out, and public confidence in data sharing and linkage more generally. Within this, there was recognition that consideration should be given to which groups of society are most important to engage with for specific projects or initiatives. There was acknowledgement, however, that there can be a lack of understanding about how to do public engagement effectively, especially among academic communities.

We also heard how, currently, agreement to data sharing by members of the public can be context-specific and dependent on who is sharing data, as well as data use. This was also reported in the Public attitudes to data and AI: Tracker survey published in March 2022 by the Centre for Data, Ethics and Innovation (CDEI). The CDEI found that organisations working on health, particularly the NHS, were most likely to be trusted by those sharing their data, with government and third sector organisations also generally preferred over private companies, where concern over data use can be greater. The purpose of data use was found to influence people's decision-making in all use-cases. This relates to another point raised with us during our interviews about miscommunication, and how it is often not made clear that data sharing for research largely involves de-identified data and that data are not going to be sold for commercial purposes. These findings and reflections demonstrate that clear and consistent communication about what is being used, and for what public benefit, are vital to gaining buy-in and avoiding a negative public response.

Finally, we heard that there is an expectation by some within the public that their data are already being shared across the public sector for the public good. This was highlighted recently in a blog by Data and Analytics Research Environments UK (DARE UK) (see Box 3) which explores the growing evidence that people want and expect data to be used for good when it is done securely and transparently. DARE UK reference their public dialogue work as evidence of this, and further examples of this type of work can be found in the case studies below. This growing evidence shows that by not sharing data, the government may be doing the opposite of achieving public good in the eyes of the public themselves – at least in certain use cases. This was also demonstrated in our own research. To understand more about public views on how public good can be served by data for research and statistics, OSR collaborated with Administrative Data Research UK (ADR UK) to carry out qualitative public dialogues with members of the public across the UK. The findings from the 68 people who participated showed strong support for data sharing, provided that best practice safeguarding is used, and participants had concerns that the missed use of data, from not sharing data, could be harmful to the public good.

Given the need for public engagement and the remaining challenges outlined above, we welcome the work of the Public Engagement in Data Research Initiative (PEDRI). PEDRI is a new sector-wide partnership looking to bring together organisations who

work with data and statistics to collaborate and embed meaningful public involvement across the data ecosystem. One of its first areas of focus is to embed best practice guidance and principles for public involvement and engagement that are specific and fit for purpose for those working in data research and statistics. This initiative could strengthen the public engagement landscape, sitting alongside other existing centres/initiatives that already support specific communities. These include the National Co-ordinating Centre for Public Engagement, which provides guidance for universities on how to plan, fund and deliver public engagement activities, and the work being taken forward by Department for Health and Social Care (DHSC), outlined in its Data Saves Lives policy paper, to "develop a standard for public engagement, setting out best practice for health and care organisations, and any other body using NHS data, to engage appropriately with the public and staff across the system on data programmes and issues".

The case studies below illustrate examples of where public engagement is being done well within the public sector and how it can inform greater understanding of social licence.

Information Box 3: Data and Analytics Research Environments UK (DARE UK)

DARE UK is a programme which aims to design and deliver a coordinated and trustworthy national data research infrastructure to support cross-topic linkage and analysis for public good. DARE UK is funded by UK Research and Innovation and puts public engagement at the heart of its work.

Case Studies: Public engagement for the public good

Secure Anonymised Information Linkage (SAIL) Databank

About: SAIL Databank is a trusted research environment (TRE) that enables research communities to access, link and analyse routinely collected population and health data within a safe and secure remote access environment.

Public engagement: SAIL Databank make regular use of their Consumer Panel which is made up of members of the public. All users of SAIL Databank can access the Consumer Panel to explore any questions they have, including feedback on research ideas, views on data protection issues and ideas for presenting findings to a public audience. Crucially, the public are involved from design to output stage of a project process for projects that are likely to have high public interest. One such project is the evaluation of the Carmarthenshire social housing initiative. This was a complex longitudinal study of the impact of improving social housing on health outcomes conducted in collaboration with Carmarthenshire County Council and members of the tenants association. Members of the SAIL Consumer Panel and local tenants associations were recruited at the beginning of the research, helped with its design and implementation and were involved in the dissemination of the results to local and national groups, including sharing a stage with Wales' First Minister.

When assessing research proposals, Research Ethics Committees (REC) look favourably upon proposals that incorporate public involvement at an early stage as this is good evidence that the research is ethically sound and in the public interest.

Thames Valley Together Project

About: The Thames Valley Violent Reduction Unit (TVVRU) was created in 2019, with funding from the Home Office (HO), to tackle the root cause of serious violence across the Thames Valley region through earlier intervention and prevention. A priority focus is the development of the first multi-agency data-sharing and analytical platform, called Thames Valley Together. It is a collaborative solution with a focus on data sharing across local authorities, health, education, policing and third sector organisations to enable response to risk factors at an individual and population level.

Public engagement: The TVVRU team have done great community engagement with young people across Oxfordshire. They held a first data ethics deliberative forum in November 2022 to consult young people on whether data should be used and, if so, how it could be used to make earlier interventions to prevent violence. The students had to consider the pros and cons of using data – thinking about issues such as privacy versus safeguarding, the need to support the most vulnerable, consent and why different agencies may need to share information with each other.

Case Studies: Public engagement for the public good

Better Outcomes through Linked Data (BOLD) Programme

About: BOLD, led by the Ministry of Justice (MoJ), is a three-year cross-government data-linking programme which aims to improve the connectedness of government data in England and Wales. It was created to demonstrate how people with complex needs can be better supported by linking and improving the government data held on them in a safe and secure way.

Public engagement: The MoJ's BOLD programme partnered with the Centre for Data Ethics and Innovation (CDEI), to undertake extensive engagement with affected groups. This included focus groups with people with complex needs, to inform the development and governance of the programme. The full report on this work can be found here.

During this work they found that participants had concerns about how data linking will be done safely and appropriately but this eased when anonymisation of personal data was explained. Participants also wanted clarity around areas such as consent, and assurances that their data will be kept safe and anonymous in the future. Both findings demonstrate the importance of transparency and of providing a thorough explanation to data subjects of what is being done with their data and exactly what data will be used.

The same research also found that participants could see how data sharing could improve public services, which they felt was a worthwhile aim.

Recommendation 1: Social Licence: The government needs to be aware of the public's views on data sharing and linkage, and to understand existing or emerging concerns. Public surveys such as the 'Public attitudes to data and AI: Tracker survey' by the Centre for Data, Ethics and Innovation (CDEI) provide valuable insight. They should be maintained and enhanced, for example to include data linking.

Recommendation 2: Guidelines and Support: When teams or organisations are undertaking data sharing and linkage projects, there is a growing practice of engaging with members of the public to help identify concerns, risks and benefits. To help teams or organisations who are undertaking public engagement work, best practice guidelines should be produced, and support made available to help plan and coordinate work. This should be produced collaboratively by organisations with experience of this work for different types of data and use cases and brought together under one partnership for ease of use. We consider that, given its current aims, the Public Engagement in Data Research Initiative (PEDRI) could be well placed to play this role.

Data security

As shown in the public engagement for the public good case studies in the previous section, ensuring and demonstrating data security is important to gaining social licence for data sharing and linkage. When we spoke to organisations and individuals working to ensure data security three topics were prominent in the discussions: trusted research environments (TREs), the Five Safes Framework and Privacy Enhancing Technologies (PETs).

Currently, a common way to make data accessible for sharing is to put it into, and make it accessible from, one or more TREs. TREs are highly secure and controlled computing environments that allow accredited researchers access to data securely stored on their servers. Government data may be held in one, none or many of these. Given data within TREs are only accessible to accredited researchers and the data flows in and out are strictly controlled, the level of security within a TRE is high. DARE UK is currently working to establish the next generation of TREs, which aim to enable fast, safe and efficient sharing, linkage and advanced analysis of data. The Integrated Data Service (IDS) is another example of a TRE currently in development across government.

The Five Safes Framework is a set of principles employed by data services, such as TREs, that enable them to provide safe research access to data. The principles of safe data, safe projects, safe people, safe settings and safe outputs are voluntarily adopted by most of the TREs and this Framework was highly praised by most we spoke to as an effective tool for ensuring the security of a data service. We did, however, hear the view that, since the Framework was developed twenty years ago, assurance that it is still able to deliver the appropriate level of security would be welcome, considering the new technologies being used to share and link data, and the increased complexity of data linkage that is occurring. There was also acknowledgement that use of the Framework is self-regulated by organisations employing it, with no overall regulator, which was concerning for some.

PETs are newer technologies that can help organisations share and use people's data responsibly, lawfully and securely. This could be by minimising the amount of data used, or by encrypting or anonymising personal information. A recent report on PETs published by the Royal Society in conjunction with the Alan Turing Institute, identifies steps to realise their benefits and their role within collaborative analysis and data governance. It describes PETs as "an emerging set of technologies and approaches that enable the derivation of useful results from data without providing full access to the data". Synthetic data are one example of a PET. Synthetic data are data created from the original data but changed in a way that preserves the characteristics of the original data while protecting the personal or sensitive information present within them.

There is growing interest in PETs and the potential benefits their use across government (and internationally) could bring. Together with the US, the UK (led by CDEI and Innovate UK) has recently announced winners of the first PETs Prize Challenges; the challenges inspired innovators in the UK and the US to build solutions that enable the collaborative development of artificial intelligence (AI) models while keeping sensitive information private. The United Nations has also established the UN

PET Lab, a collaboration of National Statistical Offices (including the Office for National Statistics (ONS), represented by the Data Science Campus) and technology experts exploring how PETs can make fully compliant data sharing between international organisations possible. In November 2022, the ONS Data Science Campus came in the top three at a UN Pet Lab hackathon that was devised to increase awareness of PETs and their potential for use by organisations to allow data access for tackling important societal and economic questions.

In his recent Pro-innovation Regulation of Technologies Review: Digital Technologies, Sir Patrick Vallance, the then Government Chief Scientific Adviser, recommends that "Government should also consider the potential use of other privacy enhancing technologies or data intermediaries as low risk options for the exchange of data…"; OSR supports this recommendation.

Recommendation 3: The Five Safes Framework: Since the Five Safes Framework was developed twenty years ago, new technologies to share and link data have been introduced and data linkage of increased complexity is occurring. As the Five Safes Framework is so widely used across data access platforms, we recommend that UK Statistics Authority review the framework to consider whether there are any elements or supporting material that could be usefully updated.

Recommendation 4: Privacy Enhancing Technologies: To enable wider sharing of data in a secure way, government should continue to explore the potential for Privacy Enhancing Technologies (PETs) to be used to enhance security and protect privacy where data are personally identifiable. The Office for National Statistics (ONS) Data Science Campus is well placed to lead and coordinate this work.

People

At every step of the pathway to share and link data, the people involved, and their skills and expertise, are instrumental to determining whether projects succeed or fail. We heard examples of departmental barriers becoming unblocked when new people arrive showing how many can be overcome simply with a new motivation, knowledge or skill. For this reason, the topics raised in this section influence many of the other topics discussed.

Leadership

We found that the biggest barrier to data sharing and linkage for some organisations is whether it is a priority for the Accounting Officer[3]. The priorities outlined for an organisation by its Accounting Officer are extremely influential as they determine the priorities of other leaders within the organisation and those responsible for enacting

[3] Within the UK government, every department has an Accounting Officer who is responsible for its day to day running as well as the department's budget. For most departments, this role is taken by the Permanent Secretary, but this is not the case for non-ministerial departments or non-government organisations. For this reason, we will refer to the most senior member of an organisation as the 'Accounting Officer'.

data sharing. We heard that different Accounting Officers have different risk appetites for data sharing, which feed into these priorities. Risk appetite can be influenced by the difference between the potential benefits and potential costs of data sharing and linkage, while the benefits may be diffuse, if something goes wrong, the effects can be very close to hand and potentially very difficult for individual organisations. This links to the view we heard that very senior leaders are more likely to focus on the risks associated with sharing data, rather than the risks of not sharing data, and points to the need for a more centralised assurance approach, to help overcome reservations of individual agents.

As well as Accounting Officers, we also heard that other people within organisations who are responsible for data access, such as data owners, can have varying levels of understanding of the challenges and opportunities associated with making more data available and accessible via data linkage. This can hamper efforts made from those working at other levels to convince senior leaders that data sharing is in the public good and make the process a lottery depending on the prior experience of those in leadership.

Those in very senior data and analytical roles across government, such as Chief Data Officers[4] and Directors of Analysis, have a big role here in championing the public benefit of sharing and linking data to their Accounting Officers and other senior leaders, and in identifying areas of success both within and outside their organisations, which can be used to demonstrate feasibility. Chief Data Officers have a particular responsibility to fulfil this role and take decisions in a way that finds balance between maintaining a focus on data security and not urging unnecessary caution. Data Protection Officers[5] can support them in this.

Making secure data sharing and linkage a strategic priority at the level of the Accounting Officer in more organisations would enable better joined up approaches across government. For this to happen, an appreciation of the potential benefits of data sharing and linkage for the public good needs to be more widely held across Accounting Officers. Sir Patrick Vallance, previously the Government Chief Scientific Adviser, provides an example of what strong leadership can look like: in his review Pro-innovation Regulation of Technologies Review he urges leaders to “prioritise wider data sharing and linkage across the public sector, to help deliver the government’s public services transformation programme.” It would also be useful to have a clear arbitration process to help resolve differences in opinion between organisations about whether data can or should be shared, due to differences in risk appetite, priorities or understanding.

[4] A Chief Data Officer (CDO) is normally responsible for organisation-wide governance and use of information as an asset.

[5] Data Protection Officers (DPOs) assist public authorities or bodies to monitor internal compliance, inform and advise on data protection obligations, provide advice regarding Data Protection Impact Assessments (DPIAs) and act as a contact point for data subjects and the Information Commissioner’s Office (ICO).

Recommendation 5: Data Literacy in Government: To gain the skills to create and support a data-aware culture, it is important for senior leaders to have awareness of and exposure to data issues. One way to raise awareness and exposure would be for senior leaders to ensure that they participate in the Data Masterclass delivered by the ONS Data Science Campus in partnership with the 10 Downing Street (No10) Data Science Team.

Recommendation 6: Data Masterclass Content: The Data Masterclass could expand its topics to include sections specifically on awareness of data linkage methodologies, the benefits of data sharing and linkage and awareness of different forms of data. This would fit well under the Masterclass topics of 'Communicating compelling narratives through data' or 'Data-driven decision-making and policy-making'.

Recommendation 7: Arbitration Process: To facilitate greater data sharing among organisations within government, a clear arbitration process, potentially involving ministers, should be developed for situations in which organisations cannot agree on whether data shares can or should occur. Developing such an arbitration process could be taken on by the Cabinet Office, commissioned by the Cabinet Secretary and delivered working with partners such as No10 and ONS[6].

Skills, knowledge and recognition

Across the UK, there is a huge demand for data roles such as data engineers and data analysts, not just in the public sector where the National Data Strategy is promoting a world-leading data economy, but also in the private sector. Recruiting people with the skills needed to link, maintain and analyse data was a significant challenge raised by many of our interviewees. Demand is outstripping supply of data skills in the UK and this is seen to be worse in the public sector as pay is often not as attractive as the private sector.

As well as recruitment, there is also a problem with retention. Retention is a particular problem for data linkage as specialist knowledge of a dataset is often held by one or two individuals, which then takes time for new staff members to learn. We have heard that staff regularly move between government departments for the opportunity of better pay as civil service pay scales differ from one department to the next for the same grade. This is exacerbated by the uptake of additional pay rewards for certain roles, such as the Digital, Data and Technology (DDaT) Pay Approach, by some departments but not others. This can cause further pay inequality across government as it allows some departments to pay bonuses on top of standard pay. Pay is not the only reason for retention issues, however, we heard that career development in data roles is not always prioritised within government, which can force those wanting to build their career to leave government altogether. We are aware of at least two career

[6] The Digital Economy Act amended the Statistics and Registration Service Act 2007 to provide the UK Statistics Authority (and ONS as its executive office) with greater and easier access to data held within the public and private sectors to support the statutory functions of the Statistics Authority in the production of official statistics and statistical research. As such, ONS is well placed to help deliver this recommendation.

frameworks for data roles within government, the DDaT career framework and the Government Analysis Function career framework. Both frameworks list what is necessary for a data role, but it is not clear how the two align and there is a lack of consistency in their use across government. This may make it hard for those working in data roles to know what skills to focus on for their development. This can be further complicated when people are members of other analytical professions as well.

The following example brings home the importance of having and retaining specialist data knowledge for the success of data sharing and linkage programmes. We heard from the UK Longitudinal Linkage Collaboration that their specialist TRE has been successful specifically because it is a collaboration of both those in the longitudinal data community and infrastructure experts. What was emphasised the most was that the specialist knowledge of the community was crucial to the development of the TRE and it would not have been as successful had it not involved industry professionals.

Recommendation 8: Career Frameworks: To enable more effective and visible support for the careers of people who work on data sharing and linkage, those responsible for existing career frameworks under which these roles can sit, such as the Digital Data and Technology (DDaT) career framework and the Analytical Career Framework, should ensure skills that relate to data and data linkage are consistently reflected. They should also stay engaged with analysts and professionals across government to ensure the frameworks are fit for purpose. These frameworks should be used when advertising for data and analytical roles and adopted consistently so that career progression is clear.

Processes

When an external researcher or government analyst wishes to access data that are held internally by government there are several high-level steps they naturally follow. Firstly, they must know the data they wish to access and where they are held. Secondly, they must establish the legal route to the data depending on the level of access required (e.g. identifiable vs non-identifiable data) and, finally, they must gain access to that data, sometimes through a TRE or possibly another route. They also need to secure the funding and/or resource to carry out their desired data project. We found that for each of these steps there are barriers which can cause significant delays.

Legal

The legal basis for data sharing was frequently raised during our interviews and views were polarised over whether relevant legislation is a barrier in itself, or whether misinterpretation of the legislation by some data holders creates a barrier. We were told that there is variation across government over how much data holders and researchers understand the process necessary to share data under their respective legal bases. If some of those data owners granting access to data are not understanding the process this can exacerbate risk aversion.

With the introduction of the Digital Economy Act (DEA) in 2017, data sharing across the public sector was further enabled under certain circumstances. These circumstances include sharing of de-identified data to produce statistics and for research purposes that are in the 'public interest' (for the purposes of this report, we interpret public interest as having the same meaning as public good and use the terms interchangeably). For research purposes, a research project needs to show that its primary purpose fits into the broad criteria listed in the DEA Research Code of Practice and Accreditation Criteria. While there is a definition of what constitutes the public interest in the DEA Research Code of Practice, under principle 4, we heard that there can still be differences in how 'being in the public interest' is interpreted, which can lead similar projects to be treated differently by the data holders responsible for granting access.

Health came up repeatedly as an area where it can be harder to share or access data due to legal restrictions. Among those we spoke to, there was a widely held conception that the DEA currently does not cover the sharing of health data for research. Speaking with the UK Statistics Authority enabled us to clarify that the restriction is slightly more nuanced: Section 64 of the DEA provides a legal route for accredited researchers to access data held by most public authorities, but it does not enable access to data held by bodies with health service functions. Other legal bases exist that allow access to data for research and statistics purposes in specific circumstances, dependent on who holds the data, what it is going to be used for and the type of data it is. But these other gateways can also come with restrictions. For example, some routes to access health data require the research purpose to have a specific and defined health benefit. New guidance developed by the National Data Guardian in 2022, which draws on insights from the public, seeks to clarify and improve public benefit evaluations by substantiating the meaning of public benefit, where health or care data is used for secondary purposes beyond care delivery.

Although not mandatory, we heard that data sharing agreements are a popular tool between two or more organisations to help navigate the legal process. We also heard that in some cases, data sharing agreements can be restrictive by only allowing a very clearly defined use case. Although not a barrier to those with defined projects, it does create a barrier to projects that aim to explore data and uncover all public good benefits the data could offer. It is also a barrier to delivering the UK Government's National Data Strategy, which has a focus around "re-using and better co-ordinating data between civil society organisations" to "create a better understanding of societal issues".

Although there are many challenges here, there is work already being done to try to make data sharing processes quicker and easier to navigate. The Central Digital and Data Office (CDDO) told us they are looking at how they can work across government to make data sharing easier and quicker while still in line with legislation. One example is setting up Memoranda of Understanding (MOUs) between departments, which agreements for data shares could then be set-up underneath when needed. The Information Commissioner's Office (ICO) also have a code of practice on data sharing which provides helpful guidance and ADR UK have produced a helpful guide on the legal framework for using administrative data for research purposes. There are also steps being taken to open up access to health and administrative data for research

purposes. For example, in November 2022, an amendment was made to Section 261 of the 2012 Health & Social Care Act (which only applies to England) to substitute "the purposes of…" with "purposes **connected with** (a) the provision of health care or adult social care, or (b) the promotion of health" [emphasis added]. The explanatory notes for the Act clarify that this new wording is intended to put beyond doubt the Health and Social Care Information Centre's (NHS England) power to share data in connection with health care or adult social care. This includes for research for purposes which benefit or are relevant to the provision of health or adult social care and developing approaches to the delivery of health and adult social care (see note 847).

We heard that the IDS is also looking at the feasibility of using a broad agreement around data sharing so that users do not have to apply for data every time they want to use it. This will also help to streamline the application process, which is a further barrier considered under 'Processes'.

Recommendation 9: Overview of Legislation: To help researchers understand the legislation relevant to data sharing and linkage and when it is appropriate to use each one, a single organisation in each nation should produce an overview of legislation that relates to data sharing, access and linkage, which explains when different pieces of legislation are relevant and where to find more information. This organisation does not need to be expert in all legislation but to be able to point people to those that are. The Office for Statistics Regulation (OSR) will help convene those in this space to understand more about who might be best placed to take this on.

Recommendation 10: Broader use cases for data: To support re-use of data where appropriate, those creating data sharing agreements should consider whether restricting data access to a specific use case is essential or whether researchers could be allowed to explore other beneficial use cases, aiming to broaden the use case were possible.

Data access

When it comes to gaining access to data, the barriers we have heard come under three main themes – data applications, finding and involving the right people, and lack of consistency and clarity. We discuss each of these separately below.

Data applications

We heard that when applying for data through a secure data platform, such as a TRE, the process is often lengthy and overly burdensome. Researchers expressed that application feedback by TREs can be drip-fed, which makes the application process longer; waits of a year or more can occur.

Researchers outside of government also spoke about being asked questions that are not relevant to the security or overall use of the output they intend to produce, but instead relate to the specific statistical methods that will be used. Researchers expressed that it is not only difficult but often impossible to know the methods that will be used until they have explored the data and understood its properties. Speaking to the UK Statistics Authority team involved in DEA accreditation, we were told that this information is needed so that projects can be assessed against the principles and

conditions in the Research Code of Practice. However, given our discussions, we judge that it could be made clearer to researchers how questions relate to DEA requirements. The UK Statistics Authority team also made us aware that there is an exploratory route available under the DEA, which allows a researcher to apply for exploratory analysis to enable them to understand the strengths and limitations of data and inform the development of more detailed research proposals. Details of this process are in the UK Statistics Authority's Research Project Accreditation Application Guidance.

Increased use of synthetic data could allow researchers to better explore data and decide how they might want to use it, so they are in a better position to make applications for the actual data with specific use cases. Research Data Scotland is already exploring using synthetic data in this way, allowing researchers to familiarise themselves with a dataset while waiting for approvals to use real data: this can aid understanding of broad cohort sizes, deriving variables of interest, and developing code, which in turn can speed up time needed for analysis once permissions have been received.

Finally, we heard there is also a frustration from government analysts that it should be easier for them to access data owned by government. As it currently stands, for most data platforms the application process is the same for both external researchers and government analysts, even though the latter have usually undergone security checks for their role. To help with this, we were told that the IDS is working with the UK Statistics Authority on a more streamlined application process for analysts when it comes to getting access to data from the Service through the DEA.

Recommendation 11. Communication: To ensure data application processes are fit-for purpose and well understood, those overseeing accreditation and access to data held in secure environments should prioritise ongoing communication with users, data owners and the public to explain and refine the information required. Wherever possible, they should offer face-to-face or virtual discussions with those applying to access data early in the process, to ensure clarity around both the data required and the process to access it.

Case study: Efficient data access – SAIL Databank

Background: SAIL Databank is a TRE, based in Swansea University, that enables research communities to access, link and analyse population and health data within a safe and secure remote access environment.

Data access processes: For SAIL, the time between data request and data access is roughly 3 months, although it can be as little as 1.5 months. For data access within a TRE, this is considered a quick turnaround time. We spoke to SAIL in depth about their process and found there are three key enablers to this:

An efficient and consistent process

The same data application process is followed each time and those that form part of the process know what is expected of them and when. A researcher who is looking to gain access to data for a research project is first put in contact with the analyst who understands the data to help them scope their project idea. Having this contact upfront helps to remove barriers around uncertainty of the data content, and its use. Their scoping form is then taken to internal review before being taken to their Independent Governance Review Panel, which considers the project from a privacy and public interest perspective. The panel take an average of 19 days to respond and the push back rate is low due to the prior internal review process. In parallel, the researcher gets their safe researcher accreditation, so when the project is approved, they are ready to access the data through the secure environment.

The longest part of this process is the first scoping phase, especially if the researcher is not clear on what they want or need. After this, the process is neat and well-defined.

A primary focus on public interest and privacy protection

Throughout the process the focus from the internal review and the approvals panel is to ensure privacy of the data subjects and to gauge the public interest in the project. In addition, the scoping phase exists to understand what support the researcher might need and to develop the research question. There are no questions that deviate from this, such as questions around methodologies used.

Reducing the need for repetitive tasks

We heard two examples of this:

1. To speed up the process of getting sign-off from data owners, SAIL have a pre-approved list of uses from the data owners. This means that if a project falls within one of the pre-approved categories it does not need to go to the data owner and potentially await their organisation's approval process.
2. Separate processes are joined up wherever possible. For example, SAIL researchers can apply for DEA data and if so, the process is split internally by SAIL so that it goes to two panels. The decision from these two panels then come together internally, so there no need for separate applications.

Finding and involving the right people

Once it is determined that data can and should be shared it is vital that the right people are involved upfront and that those people can be identified and prioritise the process needed.

For every data share there will be many teams involved such as analytical, ethical and technical teams and these can be within the same organisation or from many different ones. We have heard that not getting these teams together at the very start can cause major delays to data sharing as each team will need the time to deliver their role. It also may be counterproductive – for example, analytical teams may start to shape a dataset only to find that it includes data that either cannot be shared or are technically difficult to share.

When researchers have a question about a dataset or process it can be a challenge to find the right person within a department or team who can help. We heard how researchers can be passed between statistical and research teams without much consistency or process, which has been described to us as extremely frustrating.

We also found that there is uncertainty around data ownership and where Information Asset Owners (IAO) sit within a department. There is no national requirement for the IAO to be a certain seniority level or to sit in a certain topic area, so even when a researcher knows what data they need, finding the name and contact details of the owner can be very difficult, especially for external researchers. Taking this one step further, those needing access to multiple datasets across government expressed that it would be easier to co-ordinate data ownership at the government level with one body to oversee the process. This was a common suggestion but there are still several barriers to overcome, and it would be difficult to achieve in the short term. We hope this would be a longer-term solution that could be achieved once it is easier to navigate at the organisation level. There has been some progress already as the IDS team have set up a task and finish group to understand the challenges other organisations face when sharing data and are looking to set-up a single data sharing model.

Finally, when the right people are in place, they need to be engaged and proactive, which relates to their risk appetite, prioritisation and the time they have available. Not getting engagement was a common barrier raised, and we acknowledge it is interlinked with other barriers such as resources, leadership and understanding of the legislation.

Recommendation 12: Checklists: To ensure all necessary teams are involved at the outset of a data sharing and linking project, organisations should consider the use of a checklist for those initiating data sharing. The checklist should contain all contacts and teams within their organisation who need to be consulted to avoid last minute delays.

Lack of consistency and clarity

As alluded to in the first two themes, the process of gaining access to data is made more complex by there being different processes for different organisations and data access platforms as well as different organisational set-ups. It would be unrealistic to

recommend that all these processes become consistent with each other, but it is currently very difficult for both government and non-government researchers to know how to approach data access. In addition to this, we have heard that some departments are not aware of their own processes for data access or if any process exists. This means that each time that organisation is asked for data a person can be sent down a different route or the request is denied due to lack of knowledge on where to go or who to turn to for a decision. This lack of process is time consuming and the road to finding the correct information is repetitive. This issue has also been highlighted by CDDO in its Data Sharing Governance Framework. The second principle in the framework, named 'Make it easy to start data sharing' talks about creating a point of contact in an organisation to triage requests and queries for data sharing and access. We support this, particularly the ambition to make the point of contact easy to locate and/or to have a generic email address published on the organisation's website. CDDO, in conjunction with Government Digital Service (GDS), is also developing the Data Marketplace, which aims to provide a central place for government officials to find and understand how to access data held in other parts of government that underpin government services. Within the Data Marketplace, users will be able to 'discover' data via a Government Data Catalogue, helping to improve the discoverability of data held within government.

Recommendation 13: Transparency: Every organisation within government should be transparent about how the data they hold can be accessed and the process to follow. This guidance should be presented clearly and be available in the public domain with a support inbox or service for questions relating to the process.

Resource

When it comes to resource, we found there are two big barriers: funding and people. They are closely linked as without funding it can be impossible for staff to get the time to work on the data they need. But without upfront staffing commitments, it can be hard to show benefit and feasibility for funding to be granted.

Almost all sharing and linkage projects are collaborations between two or more government departments or external bodies. Funding structures across government are set-up so that each department controls its own spend making successful funding highly dependent on having aligned priorities and vision within each department. This siloed approach to funding ultimately affects all teams involved in the sharing and linkage process and can result in the process breaking down if just one team is unable or unwilling to get the backing needed. This is even more pronounced on projects that require sustained funding of two years or more. Spending review cycles are often tight and have strict requirements where tangible benefit needs to be shown at every decision point. For projects which are complex or require many different datasets, it may not always be possible to show benefit or meet the deadlines involved.

This siloed approach is hampering efforts of collaboration and is a primary reason why projects with external funders are often much more successful – as seen in our case study example about Administrative Data Research UK (ADR UK) and the Ministry of Justice (MoJ)'s Data First programme below. One such fund, which has helped break this siloed approach is the Shared Outcomes Fund funded by HM Treasury. The fund

is available to support pilot projects to test innovative ways of working across the public sector and government. In the 2019 spending round the fund supported a range of projects, including on drug enforcement and treatment, online harms and improving early years experiences, and all projects include collaboration and data from many organisations. The success of these projects show how sustained and ring-fenced funding can overcome barriers to data sharing and linkage.

It is worth noting that, for the third round of funding from the Shared Outcomes Fund, there will be a funding condition placed on bids. This condition will stipulate that, should an initiative seek to share data or seek funding for an analytical platform, they will have to contact and seek to partner with the IDS to achieve their goals. This is a big confidence vote in the IDS and it will need to be ready to respond to bid requests to avoid a further process barrier.

Case study: Effective funding – ADR UK and the MoJ Data First programme

Background: ADR UK is a UK data partnership with a mission to transform the way researchers access the UK's wealth of public sector data. The MoJ Data First Programme is an ambitious project with the aim of unlocking the insight stored within administrative datasets across the justice system.

Funding consideration: Recognising the potential benefits of linking data from across the justice system, MoJ contacted ADR UK to help unlock the potential of over 50 administrative datasets from across the justice system and make them accessible to accredited researchers in a secure and responsible way. Doing so would help answer questions that have immense public good implications such as: 'Are there individuals in the criminal courts who are also present in the family courts?'

Although MoJ tried to fund this work through its normal spending review cycle, the cycle was incredibly tight, and it needed to have something impactful to show at every decision point. This was not possible due to the size and scale of the data so instead MoJ applied to ADR UK for long-term, ring-fenced funding.

Ring fencing the funding in this way was crucial to success and meant external factors such as the pandemic and cost of living crisis did not disrupt progress. There was also an academic embedded in the team, which helped bring the knowledge of the research community into the development. It is important to note that prior to approaching ADR UK, MoJ had support from its leadership team, a motivation for public good outcomes and the drive needed to succeed in its partnership with ADR UK.

The Data First Programme has produced some high impact datasets that allow researchers to understand the extent and nature of repeat users of the magistrates and Crown courts including the type of offences committed and to explore how children's education and social care factors in England relate to offending behaviours.

Recommendation 14. Funding Structure: To allow every organisation a consistent funding stream for their projects, a centralised government funding structure for data collaboration projects across government, such as the Shared Outcome Fund, should be maintained and expanded.

Technical

Finally, we discuss the technical elements of data and data linkage that were raised during our review. We heard it can be a real challenge for those linking data to get enough information about the data they are working with to provide a high-quality linked output with a measurable rate of error. Regarding data linkage methodologies, while we heard many positive reflections on the effectiveness of current methodologies, and the way that these are being developed, it was also acknowledged that methodological challenges do still exist, which can also themselves lead to issues with the quality of linked data. Finally, we heard how variation in the data standards and definitions used across government is making linking harder. These three areas are discussed in more detail below.

The quality of metadata

Metadata is a set of information which describes data. When using data for linkage purposes it is important to have access to as much metadata as possible as it contains information about where the data have come from and how they were collected. This is important because linking data often relies on matching cases by making assumptions about the data based on characteristics, in a process also known as 'fuzzy matching'.

We heard that data held within government, at the level of both the dataset and the data descriptors, are not well documented making it difficult for a researcher to know if a project is feasible. A typical linkage research project requires the following:

1. A research question of interest
2. Understanding of what data would be needed to answer the research question
3. Knowledge of whether the right data exists (description of dataset, description of the variables, coverage of the data)
4. Knowledge of who owns the data (department and owner)
5. Knowledge of how those variables were collected for linkage

Stages three, four and five are where a lack of metadata can cause significant problems to a data linkage project.

Firstly, researchers stated that they often do not know what data are held within the government. They described how it can be impossible to know if a dataset exists and if so, if the variables contained within it will be useful to them. We have heard examples of how both academics and government analysts have resorted to applying for data based on a brief description on the off chance that it might contain the variables they need. When they receive the data they then find that there is no further information provided about what the data are to help understand how it can be linked with other

data. Given these data are then un-usable this has wasted both their time, and the time of everyone involved in the data access process.

Secondly, when a lack of documentation exists, people need to resort to the knowledge of people who work on or own the data. The difficulty of finding data owners within a department has been mentioned previously but academics have expressed how it is also difficult to know in which department the data might sit. Without the information contained in the metadata or being able to contact someone who knows about the data, researchers can be left unable to conduct their research.

There are also considerations around how not having this metadata affects the quality of a linked output. We heard a lot about quality when talking to those who work in specialist linkage teams within government. The main concern was that a lack of information about how data were collected and processed and the limitations they contain prior to conducting linkage would undoubtedly cause errors and inconsistencies in the linkage process. These errors can then be exacerbated by the linking of linked datasets. Quality assurance is also affected as it is hard to understand how the linking has been done which makes reproducibility impossible.

Lack of metadata is not usually a result of poor upfront planning but because most data were not originally collected for the purpose of linking and therefore it was not considered a necessity. We heard that some departments are better at providing metadata than others, and this is usually because they have been allocated resource to make their data more understandable. There is optimism that moving more data to platforms that are DEA accredited will encourage more departments to think about improving their metadata.

There is work being done in this area to help overcome this barrier to data linkage. As part of the Data Marketplace previously mentioned, CDDO is developing a Metadata Requirements Specification for government, which builds on existing guidance on metadata standards and the IDS is working closely with CDDO to set-up a metadata standard model, so that all government organisations contributing to the platform can follow one standard. ADR UK have also developed a data catalogue to help with the discoverability of data that has been made available for public good research. The catalogue contains information on the department which holds the data, a description of the data and, in some cases, the data dictionary of variable descriptions. This catalogue brings together metadata from the four nations of the UK, which was only previously available by searching four separate catalogues. This catalogue is publicly accessible and should help researchers overcome some of the barriers described above, particularly around knowing what data exists within government.

Recommendation 15. Sufficient resources: To enable effective, efficient, and good quality data linking across government, senior leaders should ensure there are sufficient resources allocated to developing quality metadata and documentation for data held within their organisations.

Data linkage methodologies

Throughout our interviews we heard that there are lots of different data linking methodologies currently in use across government, each of which have their own strengths and limitations with respect to the quality of the linked data they produce. We in OSR are not technical experts in data linkage, and therefore we are not best placed to make judgements on what the 'correct' or 'best' ways to link data are. As the regulator for Official Statistics, we advocate for ensuring these methods are developed and deployed in a way that supports public confidence in them and any resulting analysis or research. This means a focus on transparency, openness, and collaboration. Teams and support groups working across government to support data linkage have a big role to play in enabling improvements in linkage methodologies and their application. The National Statistician's guidance on joined up data in government, which also highlighted the need for the data linkage community to work together, provides links to community groups that could support those working on linkage projects across government, as well as a series of peer-reviewed articles from academia, government and the third sector on linkage methodologies. We also heard about the Government Data Architecture community, which exists to share ideas, experiences and methods, in an effort to standardise the way organisations work with data and ease communication between government departments.

Data standards and definitions

Data standards are accepted agreements on the format, structure, definition, and manipulation of data. When everyone follows the same data standards can make it easier to share data securely, to understand how data can be linked and for departments to automate their processes for data linkage, which saves resource.

Data standards in government are currently not consistent across departments or within departments over time. We heard how this lack of a 'common language' and awareness of how other departments manage their data is causing repetition of work which could have been standardised. This also highlights the issues of poorly aligned systems whereby non-standardised data encourages the use of many different systems, which make data exchange more difficult. Currently, this barrier is mainly a problem to those working on large cross-government linkage projects as this is where the most time is lost to repetition. However, many of those we spoke to expressed concern that this could pose a big challenge to the collaborative and accessible future for data sharing and linking that is envisaged by the UK government.

There is work being done across government to help align data standards. Mission 1 of the National Data Strategy has a priority to "promote the development and use of good data standards so that data is held, processed and shared according to the FAIR principles" and the Department for Science, Innovation and Technology is mapping current data standards across government. The GDS is working to develop the Government Data Exchange, which aims to provide infrastructure for sharing data between government departments. Finally, the Data Standards Authority, led by the CDDO, is working to improve data standards across central government.

Recommendation 16: Standardisation: Many departments are looking to standardise government data and definitions, but it is unclear whether or how these

initiatives are working together. Those working to standardise the adoption of consistent data standards across government should come together to agree, in as much as is possible for the data in question, one approach to standardisation which is clear and transparent. Given the work done by the Data Standards Authority, led by the Central Digital and Data Office (CDDO), the CDDO may be best placed to bring this work together.

Chapter 2 – The future of data sharing and linkage across government

In Chapter 2 we will look to the future of data sharing and linkage in government, helping bring to life the barriers and enablers presented in Chapter 1. We present four possible 'future scenarios' for data sharing and linkage, set five years from now, based around the themes raised in our interviews. Future scenarios are not predictions but stylised versions of possible futures. We believe these help to bring out the impact on public good of acting on (or not acting on) the current barriers that exist to data sharing and linkage. They allow the reader to explore the possible implications of their choices when making decisions in this space. The four scenarios we consider are: Data Sharing and Linkage for Public Good, Data Sharing and Linkage in Silos, Data Sharing and Linkage for Government and Data Sharing and Linkage Deprioritised.

To support and illustrate the scenarios, we have developed three 'personas', which outline the potential experiences of an academic researcher, a government researcher, and a service coordinator working in the charity sector. These emphasise the impacts and outcomes of different scenarios and illustrate the argument for making choices that lead towards data sharing and linkage for the public good.

Finally, we present our 'roadmap' to the scenario: Data Sharing and Linkage for Public Good. This roadmap is informed by the discussions presented in Chapter 1. It highlights where the current data sharing and linkage landscape across government is now, where we would like it to see it go, and the recommendations we have made that will help to get there.

Four alternative futures

To keep the scenarios consistent with each other, each scenario has the same four themes running through them, as discussed in Chapter 1. These are:

- Public engagement and social licence: The importance of obtaining a social licence for data sharing and linking and how public engagement can help build understanding of whether/how much social licence exists and how it could be strengthened. We also explore the role data security plays here.
- People: The risk appetite and leadership of key decision makers and the skills and availability of staff.
- Processes: The non-technical processes that govern how data sharing and linkage happens across government.
- Technical: The technical specifics of datasets, as well as the infrastructure to support data sharing and linkage.

Scenario 1: Data Sharing and Linkage for Public Good

In this scenario, public understanding and buy-in to the benefit of data being shared and linked is high. Different groups across society can see the positive outcomes and the cultural norm is to be trusting, pro-collaborative and engaged with data that affects them. Furthermore, the outcomes of research using linked data are transparently published and widely accessible to all, leading to a willingness among members of the

public to allow their data to be shared and used for public good. Public confidence is supported by consistent demonstration from those sharing and linking data that security and privacy are high priority. Where the data are personally identifiable, Privacy Enhancing Technologies (PETs) are used to enhance security and protect privacy.

Strong partnerships exist within and between government organisations, and extend beyond this to include external researchers, partnership organisations, the wider public sector and the private sector. Senior leaders understand and champion the benefits of sharing and linking data, actively encouraging and promoting safe and secure research using linked data for the public good by promoting a can-do culture and being proactive in removing barriers. Staff feel valued and supported which has created a trusting and collaborative environment across government leading to greater staff retention.

Access to government data is consistent and streamlined, making it more transparent and easier for those both in and beyond government to find and engage with the data they need. Both the data and metadata are of high quality and are provided 'linkage ready', where appropriate, reducing the time it takes researchers to provide public good research and reducing the time burden on analysts. Funding is effectively prioritised and sufficiently maintained to allow far and wide-reaching impacts at both local and national levels.

Opportunities to enhance the public good of data and statistics are fully realised and missed data use is very rare.

Scenario 2: Data Sharing and Linkage in Silos

In this scenario, data sharing and linking is happening in silos across government, usually aided by partnership organisations. Public understanding about what is happening with data and what public good impact it is having is confused and even though some groups in society are grateful for the areas where engagement and transparency have been good, other groups are frustrated that more is not being done in specific areas. This confusion is leading to reservation for some when considering willingness to share data, even in areas that have good engagement due to the lack of clarity from government as a whole.

In the silos where good progress is happening, senior leaders are proactive and engaged, collaboration is high, and consistency of practices helps things run smoothly. However, this positive approach is not replicated in all areas and there are pockets where little to no progress is made.

Funding is not evenly distributed and usually goes to those who have already had success, leaving areas with high potential but disengaged leaders worse off. Staff experiences differ widely from feeling supported and driven in pockets where progress is good to feeling underutilised and frustrated where it is not. This is leading to high staff turnover between departments. Access to data is inconsistent and for researchers it is luck as to whether the data they want falls within a successful pocket of work. This is the same with data quality where some data are very well documented and structured whereas others are not.

Public good is being realised in certain topic areas, but data from other topic areas could provide a more enhanced picture and opportunities are likely being missed. The frustration and confusion among the public is undermining their trust in government and thus jeopardising government's social licence in relation to data sharing and linkage.

Scenario 3: Data Sharing and Linkage for Government

In this scenario, data are shared and linked well across government but the value and benefit to those external to government is not being considered or realised. As a result, public understanding of the government's use of data and the impact it is having on public services is limited. This is leading to a lack of willingness to share data with government and is helping misinformation to spread more easily. This, in turn, is increasing levels of mistrust and making government more vulnerable to public backlash. The ability of government to continue to share and link data is threatened due to their lack of openness and the wider impact this is having.

Within government, leaders are proactive and encouraging of sharing but only within the protected government environment, with outputs developed for internal use. As a result, government analysts find the data access process simple, consistent and streamlined and enjoy working within a high collaboration environment. Funding is also effectively distributed across government departments giving each department the incentive to make their data high quality and well-documented for other government analysts.

Outside of government the picture is very different. Academics and researchers do not have a defined or consistent pathway to data access and find it difficult to know who to talk to resolve their situation. Those that have found success have found it can take many years and research grants have expired before data have become available. Furthermore, government are not engaging with the wider public and haven't made any outputs from their analysis available in the public domain.

This scenario is good for internal government management but public good is not being realised and 'missed use' of data is common. It is also fragile and faces the risk of a rapid loss of social licence for data sharing and linkage.

Scenario 4: Data Sharing and Linkage Deprioritised

In this scenario, data sharing and linkage is not a priority for government. There is a view from senior leaders that 'something has been done' and therefore there is no incentive to go any further. As a result, public understanding of the use of data is limited and there are no measurable improvements to public services or processes being seen. This is causing an unwillingness amongst the different sections of society to share data. These sections increasingly question why data that they know is being collected is not being used in more innovative ways to improve their lives.

Vacancies are not being filled and the analysts that are still working in this area feel frustrated, un-motivated and un-supported in their specialities with no sign of this improving. Government data skills are falling dangerously behind the private sector meaning any new government data are not being processed or managed effectively.

Funding has also dried up and partnership organisations are finding it more difficult to embed their messages and practices within the departments themselves.

Although data exists and can be accessed by analysts and researchers, the amount available is limited to already existing projects and there are no formal processes for data access or linkage. This leads to a feeling of 'right place, right time' when trying to get data access and a prior knowledge of who to speak to. When data does become available it is not always clear what the data are and their structure is often unusable in their raw state. As a result, time is wasted doing the same tasks each time data access is granted. Collaboration within and beyond government has slowed and dialogue rarely happens outside of small teams. This is further isolating those trying to do projects that have public good potential.

Although there was the potential for data sharing and linkage for the public good, this has not been realised and there are many examples of missed opportunities where data could have a real impact.

Visualising the scenarios

Below are two visualisations that represent how the scenarios interrelate with one another. These have been included to show the importance of both internal collaboration and external engagement on the future public good that data sharing and linkage can provide. Put differently, both 'internal collaboration' and 'external engagement' underpin the likelihood of arriving in each scenario, which in turn has a level of public good attached to it.

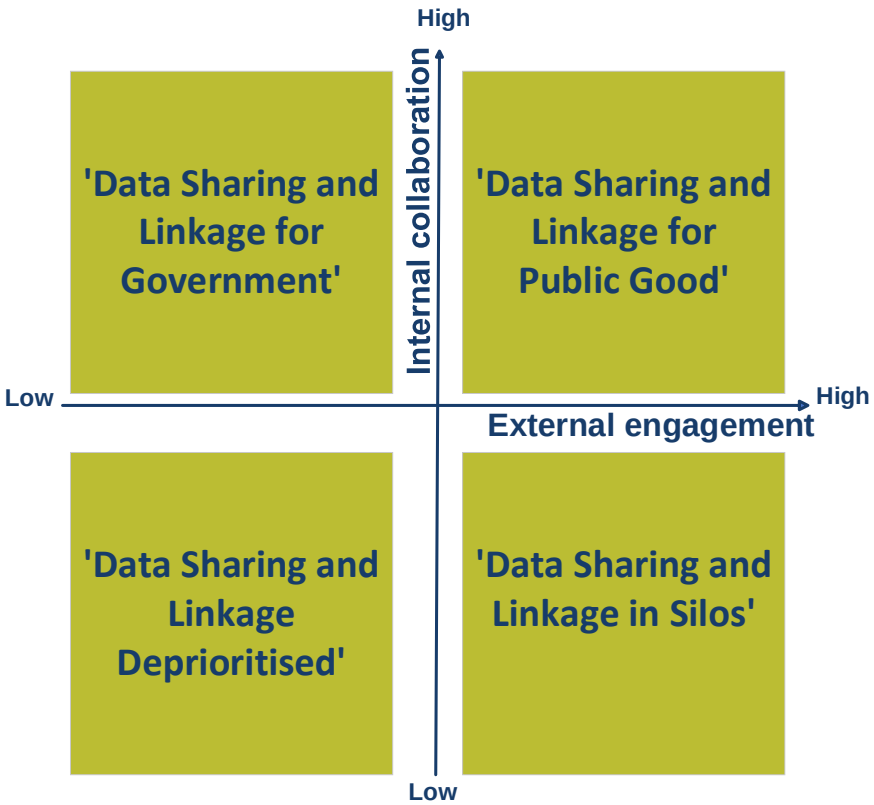

Figure 1: The four scenarios based on their level of external engagement and internal collaboration across government.

Figure 2: The four scenarios based on the level of public good achieved.

Personas

To support and illustrate the scenarios presented above, we have developed three imaginary personas: an academic researcher, a government researcher and a service coordinator working in the charity sector. For each, we have imagined their background, 'data mission' and the experience they might have in each scenario.

Academic Researcher

Name: Steve
Occupation: Professor at a university
Location: Edinburgh

Background: Steve is the head of a small team of researchers based in the social science department of a university. Their research focuses on the ways in which adverse childhood experiences impact on adult mental health. Steve is particularly interested in the links between childhood deprivation and the diagnosis of severe psychiatric disorders, such as bipolar disorder and schizophrenia. The team typically conduct their research using large, linked administrative datasets.

Data mission: Steve and his team have received funding for two years for a project which maps out indicators of childhood deprivation, such as receiving free school meals, and residing in a household in which one or more parent is in receipt of disability or incapacity benefit, with adult mental health outcomes, such as the prescription of psychiatric medications or a diagnosis of a psychiatric disorder. Steve wants to link data from the Department for Education (DfE), the Department for Work and Pensions (DWP) and the NHS.

Response to Data Sharing and Linkage for Public Good

Steve and his team provide evidence that public good can be achieved through their research and as a result, they are granted access to a linked administrative dataset through a secure data access platform. The dataset contains linked data from the DfE, DWP and NHS. This means that Steve's team receive their data in a timely manner and can complete their research within their funded period. Their work is widely used by organisations within and beyond the public sector.

Response to Data Sharing and Linkage in Silos

Although Steve's team successfully obtain permission to work with a linked dataset, they struggle to link the datasets required for them to complete their analysis. The mechanisms are not in place for data sharing between the two government departments and the health service, the result of this being that full data linkage cannot be performed during their funded period. They successfully link two of the three data sources, resulting in some outputs.

Response to Data Sharing and Linkage for Government

Steve and his team struggle to form working relationships with each of the three organisations from which they require data. They are aware of data linkage happening within government but have been unable to gain permission to use the data themselves. As a result of this, they cannot perform the data linkage within their funded period.

Response to Data Sharing and Linkage Deprioritised

Steve and his team are unable to form working relationships with any of the organisations from which they seek data. The team are also aware that data linkage is not being routinely performed within government and as a result they are not able to use a previously linked dataset. They are unable to answer their research questions in their funded period.

Government Researcher

Name: John
Occupation: Social Researcher, Ministry of Justice (MoJ)
Location: Sheffield

Background: John leads a team of researchers at the MoJ, who are working to understand the impacts of parental imprisonment on the educational outcomes of children. They would like to compare the educational outcomes of children whose parents have criminal records but without a custodial sentence with those with a parent who has been in prison.

Data mission: John and his team want to link up data held by the Department for Education (DfE) with records from HM Prison Service (HMPS), for children whose parents have been in prison, and the Police National Computer (PNC), for those whose parents have committed crimes but have not been in prison. The team are aiming to link data over a period of ten years, to enable them to understand the long-term impacts of parental imprisonment.

Response to Data Sharing and Linkage for Public Good

John and his team are successful in their attempts to link education attainment data with data from both HMPS and the PNC. They can build an anonymised, longitudinal dataset, containing data on the attainment of children whose parents have criminal convictions and whether they have served custodial sentences. There has been a high degree of public trust in the project due to the levels of transparency around the project and the amount of engagement conducted with stakeholders.

Response to Sharing in Silos

John and his team can link data from HMPS with data from the DfE, allowing them to understand the link between parental imprisonment and educational outcome. However, they are not able to link with the data from the PNC. This means that while they have a good understanding of the impacts that parental imprisonment may have on a child, they aren't able to determine whether these impacts occur because of the time their parent has spent in prison, or the criminal conviction.

Response to Data Sharing and Linkage for Government

John and his team are successful in their attempts to link all three of their datasets, which allows them to answer their research questions. They produce a report and use their findings to inform policy around families and the criminal justice system. There is however very little engagement outside of government and the public are mostly unaware that the data are being linked. The lack of public awareness of the project means that stakeholders, such as children's charities and non-government researchers, are unable to use the findings from the research.

Response to Linkage Deprioritised

John and his team are unable to link data from the MoJ with the HMPS and the DfE. Instead, they are encouraged to use a previously linked dataset, which allows them to partially answer their research questions. There is little interest from external organisations, as there is little awareness of data linkage performed by government departments.

Employee in the Charity Sector

Name: Martha
Occupation: Service Coordinator, charity sector
Location: Manchester

Background: Martha works for a small charity which helps individuals experiencing homelessness. The charity provides practical assistance for their service users, including food and short-term accommodation. They also provide advice, enabling their service users to access healthcare and benefits in the short term, and permanent housing and employment in the long term. Martha's team have recently started conducting their own research with their service users.

Data mission: Martha needs to know about the lives of those affected by homelessness. She is particularly interested in the health impacts of rough sleeping, as well as the long-term housing and employment outcomes for individuals who have previously experienced homelessness. This information will allow the charity to tailor the advice and the support they deliver to the needs of their service users.

Response to Data Sharing and Linkage for Public Good

Martha can access an abundance of information about the long-term outcomes of people affected by homelessness. She can use data from a longitudinal study on the employment outcomes for individuals who have previously experienced homelessness to inform the advice she gives to her service users, which leads to an increase in the number of service users gaining employment. The charity is considering submitting their own operational data for use in a large research project, having seen the benefits of research using linked datasets. They have confidence in the safety of the data.

Response to Data Sharing and Linkage in Silos

Martha is aware that there are some public sector research projects which use linked data. However, these projects often do not include individuals who have previously experienced or are currently experiencing homelessness, so she is unable to build complete pictures. There is little clarity around the reasons for some areas being prioritised over others, which leads to distrust, with the charity being reluctant to share data in the future.

Response to Data Sharing and Linkage for Government

Within government, research is being conducted about the longitudinal outcomes of individuals who have previously experienced homelessness. However, this research is mostly being conducted for internal use, which means that practitioners employed in the charity sector are not aware of the work and cannot use or help others benefit from the results of it. They are also disinclined to share their data, as they are not aware of previous incidences when data sharing has been of benefit.

Response to Data Sharing and Linkage Deprioritised

There is no longitudinal, linked dataset on the long-term outcomes of individuals who have previously experienced homelessness. This means that although Martha can use other sources of data to inform her practice, she does not have data about longer term outcomes, which would have been useful for her service users. The charity is also reluctant to share their data, as there are few examples in the public domain of cases of successful data linkage.

A roadmap to Data Sharing and Linkage for the Public Good

This section maps out how our recommendations can take us from where the data sharing and linkage landscape is now, within government, to where we think it should aim to be. We do this by linking our recommendations to our ideal scenario 'Data sharing and Linkage for Public Good'.

The current data sharing and linkage landscape across government	**What do we want it to look like?**	**Relevant recommendations**
Public engagement and social licence: Public engagement		
There is a need for more public engagement about data sharing and linkage, to improve both transparency of work that is being carried out, and public confidence in data sharing and linkage more generally. There is growing evidence that people in the UK want and expect data to be used when it is done securely and transparently. There is an expectation by some among the public that their data are already being shared and linked within the public sector for the public good. There are examples of where public engagement is being done well, informing greater understanding of social licence. However, there was acknowledgement that there can also be a lack of understanding about how to do public engagement effectively.	"Public understanding and buy-in to the benefit of data being shared and linked is high. Different groups across society can see the positive outcomes and the cultural norm is to be trusting, pro-collaborative and engaged with data that affects them. Furthermore, the outcomes of research using linked data are transparently published and widely accessible to all, leading to a willingness among members of the public to allow their data to be shared and used for public good."	**Recommendation 1: Social Licence:** The government needs to be aware of the public's views on data sharing and linkage, and to understand existing or emerging concerns. Public surveys such as the 'Public attitudes to data and AI: Tracker survey' by the Centre for Data, Ethics and Innovation (CDEI) provide valuable insight. They should be maintained and enhanced, for example to include data linking. **Recommendation 2: Guidelines and Support:** When teams or organisations are undertaking data sharing and linkage projects, there is a growing practice of engaging with members of the public to help identify concerns, risks and benefits. To help teams or organisations who are undertaking public engagement work, best practice guidelines should be produced, and support made available to help plan and coordinate work. This should be produced collaboratively by organisations with experience of this work for different types of data and use cases and brought together under one

		partnership for ease of use. We consider that, given its current aims, the Public Engagement in Data Research Initiative (PEDRI) could be well placed to play this role.
Public engagement and social licence: Data security		
The amount social licence for a data sharing or linkage project can be related to data security. The Five Safes Framework is a set of principles employed by data services, such as TREs, that enable them to provide safe research access to data. Assurance that it is still able to deliver the appropriate level of security would be welcome. Privacy Enhancing Technologies (PETs) are newer technologies that can help organisations share and use people's data responsibly, lawfully and securely. There is growing interest in PETs and the potential benefits their use across government (and internationally) could bring.	"Public confidence is supported by consistent demonstration from those sharing and linking data that security and privacy are high priority. Where the data are personally identifiable, Privacy Enhancing Technologies (PETs) are used to enhance security and protect privacy."	**Recommendation 3: The Five Safes Framework:** Since the Five Safes Framework was developed twenty years ago, new technologies to share and link data have been introduced and data linkage of increased complexity is occurring. As the Five Safes Framework is so widely used across data access platforms, we recommend that UK Statistics Authority review the framework to consider whether there are any elements or supporting material that could be usefully updated. **Recommendation 4: Privacy Enhancing Technologies:** To enable wider sharing of data in a secure way, government should continue to explore the potential for Privacy Enhancing Technologies (PETs) to be used to enhance security and protect privacy where data are personally identifiable. The ONS Data Science Campus is well placed to lead and coordinate this work.
People: Leadership		
Strong collaboration between the UK statistical system and ADR UK has supported linkage and sharing of administrative datasets within	"Strong partnerships exist within and between government departments, and extend beyond	We do not have a specific recommendation against this ambition, but our other

and across organisations in all four UK Nations and is helping to make them available to accredited researchers within and beyond government in a safe and secure way.	this to include external researchers, partnership organisations and the local and private sectors.”	recommendations seek to enhance collaboration across government.
At every step of the pathway to share and link data, the people involved, and their skills and expertise, are instrumental to determining whether projects succeed or fail. The biggest barrier to data sharing and linkage for some organisations is whether it is a priority for the Accounting Officer. Making secure data sharing and linkage a strategic priority at the level of the Accounting Officer in more organisations would enable better joined up approaches across government. For this to happen, an appreciation of the potential benefits of data sharing and linkage for the public good needs to be more widely held across Accounting Officers.	“Senior leaders understand and champion the benefits of sharing and linking data, actively encouraging and promoting safe and secure research using linked data for the public good by promoting a can-do culture and being proactive in removing barriers.”	**Recommendation 5: Data Literacy in Government:** To gain the skills to create and support a data-aware culture, it is important for senior leaders to have awareness of and exposure to data issues. One way to raise awareness and exposure would be for senior leaders to ensure that they participate in the Data Masterclass delivered by the Office for National Statistics (ONS) Data Science Campus in partnership with the 10 Downing Street (No10) Data Science Team. **Recommendation 6: Data Masterclass Content:** The Data Masterclass could expand its topics to include sections specifically on awareness of data linkage methodologies, the benefits of data sharing and linkage and awareness of different forms of data. This would fit well under the Masterclass topics of ‘Communicating compelling narratives through data’ or ‘Data-driven decision-making and policymaking’. **Recommendation 7: Arbitration Process:** To facilitate greater data sharing among organisations within government, a clear arbitration process, potentially involving ministers, should be developed for situations in which organisations cannot agree on whether data shares can or should occur. Developing such an arbitration

		process could be taken on by the Cabinet Office, commissioned by the Cabinet Secretary and delivered working with partners such as No10 and ONS.
People: Skills, knowledge and recognition		
Recruiting people with the skills needed to link, maintain and analyse data was a significant challenge raised by many of our interviewees. As well as recruitment, there is also a problem with retention. We heard that staff regularly move between government departments for the opportunity of better pay as civil service pay scales differ from one department to the next for the same grade. Career development in data roles is not always prioritised within government.	"Staff feel valued and supported which has created a trusting and collaborative environment across government leading to greater staff retention."	**Recommendation 8: Career Frameworks:** To enable more effective and visible support for the careers of people who work on data sharing and linkage, those responsible for existing career frameworks under which these roles can sit, such as the Digital Data and Technology (DDaT) career framework and the Analytical Career Framework, should ensure skills that relate to data and data linkage are consistently reflected. They should also stay engaged with analysts and professionals across government to ensure the frameworks are fit for purpose. These frameworks should be used when advertising for data and analytical roles and adopted consistently so that career progression is clear.
Processes: Legal and Data access		
There is variation within government over how much data holders and researchers understand the process necessary to share data under their respective legal bases.	"Access to government data is consistent and streamlined making it more transparent and easier for those both in and beyond government to find and engage with the data they need."	**Recommendation 9: Overview of Legislation:** To help researchers understand the legislation relevant to data sharing and linkage and when it is appropriate to use each one, a single organisation in each nation should produce an overview of legislation that relates to data sharing, access and linkage, which explains when different pieces of legislation are relevant and where to find more

When applying for data through a secure data platform, the process is often lengthy and can appear overly burdensome. For every data share there will be many teams involved such as analytical, ethical and technical teams and these can be within the same organisation or from many different ones. We have heard that not getting these teams together at the very start can cause major delays to data sharing. When researchers have a question about a dataset or process it can be a challenge to find the right person within a department or team who can help. We also found that there is uncertainty around data ownership and where Information Asset Owners (IAO) sit within a department.		information. This organisation does not need to be expert in all legislation but to be able to point people to those that are. The Office for Statistics Regulation (OSR) will help convene those in this space to understand more about who might be best placed to take this on. **Recommendation 10: Broader use cases for data:** To support re-use of data where appropriate, those creating data sharing agreements should consider whether restricting data access to a specific use case is essential or whether researchers could be allowed to explore other beneficial use cases, aiming to broaden the use case were possible. **Recommendation 11: Communication:** To ensure data application processes are fit-for purpose and well understood, those overseeing accreditation and access to data held in secure environments should prioritise ongoing communication with users, data owners and the public to explain and refine the information required. Wherever possible, they should offer face-to-face or virtual discussions with those applying to access data early in the process, to ensure clarity around both the data required and the process to access it. **Recommendation 12: Checklists:** To ensure all necessary teams are involved at the outset of a data sharing and linking project, organisations should consider the use of a checklist for those

		initiating data sharing. The checklist should contain all contacts and teams within their organisation who need to be consulted to avoid last minute delays. **Recommendation 13: Transparency:** Every organisation within government should be transparent about how the data they hold can be accessed and the process to follow. This guidance should be presented clearly and be available in the public domain with a support inbox or service for questions relating to the process.
Processes: Resource		
Funding structures across government tend to be set-up so that each department controls its own spend, making successful funding highly dependent on the priorities and vision within each department. This siloed approach to funding means data sharing/linking projects are susceptible to breaking down if just one team is unable or unwilling to get the backing needed. Spending review cycles are often tight and have strict requirements where tangible benefit needs to be shown at every decision point. For projects which are complex or require many different datasets it may not always be possible to show benefit or meet the deadlines involved.	"Funding is effectively prioritised and sufficiently maintained to allow far and wide-reaching impacts at both local and national levels."	**Recommendation 14: Funding Structure:** To allow every organisation a consistent funding stream for their projects, a centralised government funding structure for data collaboration projects across government, such as the Shared Outcome Fund, should be maintained and expanded.

This siloed approach is hampering efforts of collaboration and is a primary reason why projects with external funders are often much more successful.		
Technical: The quality of meta data and Data standards and definitions		
It can be a real challenge for those linking data to get enough information about the data they are working with to provide a high-quality linked output with a measurable rate of error. Variation in data standards and definitions used across government is making linking harder.	"Both the data and metadata are of high quality and are provided 'linkage ready', where appropriate, reducing the time it takes researchers to provide public good research and reducing the time burden on analysts."	**Recommendation 15: Sufficient resources:** To enable effective, efficient, and good quality data linking across government, senior leaders should ensure there are sufficient resources allocated to developing quality metadata and documentation for data held within their organisations. **Recommendation 16: Standardisation:** Many departments are looking to standardise government data and definitions, but it is unclear whether or how these initiatives are working together. Those working to standardise the adoption of consistent data standards across government should come together to agree, in as much as is possible for the data in question, one approach to standardisation which is clear and transparent. Given the work done by the Data Standards Authority, led by the Central Digital and Data Office (CDDO), the CDDO may be best placed to bring this work together.

Next Steps

Following the publication of this report the Office for Statistics Regulation (OSR) will be reaching out to those organisations mentioned in our recommendations to discuss how they might be taken forward, including any support we, at OSR, can give. We will also consider convening cross-government round tables to discuss the themes we have outlined and to help join-up organisations who might be working on similar solutions or challenges. To help generate and maintain momentum on our recommendations, we will review and publicly report on progress towards them between six months and one year after this publication.

As part of our wider regulatory role, we will continue to focus on data sharing and linkage, influencing more joined up approaches where we can. We will continue to identify opportunities to champion the value of data sharing and linkage, including through sharing case studies that illustrate how organisations across government are making progress on data sharing and linkage for the public good. We are also considering how we might play a role in independently reviewing wide reaching data sharing initiatives, such as the Integrated Data Service, to help give users assurances of their trustworthiness, quality and value. Further information regarding our work in this space will be announced on our website.

Annex A: Method

Research questions / purpose

The over-arching aim of the Office for Statistics Regulation's work was to understand how the data sharing and linkage landscape across government has changed since our previous work in this area. Specifically, we wanted to know what progress has been made since our last reports, and what barriers to effective data sharing and linkage still exist. We also wanted to identify examples of good practice and understand the enablers for good work that is taking place.

Participants

We spoke to a range of stakeholders from government, the wider public sector, and from the private sector representing a range of roles and responsibilities in relation to data sharing and linkage. These included statisticians and analysts working on projects using linked datasets, those actively involved in linking data, those responsible for managing and running trusted research environments (TREs), and those who facilitate data sharing and linkage, for example by funding projects.

To ensure we captured a broad range of views we spoke to both those directly involved in projects and senior leaders/those with strategic oversight of projects and programmes.

We identified participants proactively ourselves and through advertising the project on our website and asking interested parties to contact us. After the initial interviews, we also identified further participants via "snowballing", where those we had already spoken to recommended further key individuals or teams that it would be beneficial for us to engage with. We sought to engage with stakeholders from each of the four UK countries to ensure views from each administration were captured, as well as those involved in data linkage projects at a regional or local level.

We concluded our interviews when we felt that we had reached the point of data saturation, i.e. we were no longer discovering new information in our analyses of the data. At this point we continued to invite feedback and contributions via email.

To ensure representation across a wide range of views, we grouped stakeholders into four categories:

- Those using linked datasets
- Those involved in linking data
- Those providing data to be linked, but not necessarily carrying out the linkage work themselves
- Those involved in data sharing and linkage, but not necessarily doing the work directly themselves

In practice, not all stakeholders fitted neatly into just one of the four categories but having this breakdown allowed us to ensure we spoke to a broad range of people across the whole spectrum of data sharing and linkage. A list of the organisations and teams we engaged with is available in Annex B.

Approach

We carried out semi-structured interviews to explore stakeholders' experiences and perceptions of data sharing and linkage. Semi-structured interviews involve open-ended discussion with participants, guided by a pre-determined discussion plan.

We devised an interview schedule containing questions for the four different stakeholder types outlined above. The questions were broadly similar across the four categories, but the wording was altered where appropriate to make them relevant to each type of stakeholder and their specific situation. The questions focussed broadly on what work was being done in relation to data access and linkage, enablers and barriers, and visions for the future. As well as higher level questions, we also devised some prompt / follow-up questions for use where necessary.

The interviews were generally carried out virtually using Microsoft Teams. Two members of the team were present for each interview, with one member leading the conversation and the other taking notes. The interview schedule was used as a guide but generally the conversations took an open format and were allowed to flow to capture the unique experiences of each of the individuals and teams we spoke to. Interviews usually lasted in the region of forty-five minutes to one hour.

Analysis

We used a thematic analysis approach to identify and develop common themes in the interview data. Initially we started with the broad headings of barriers and facilitators; we then identified further themes and sub-themes and coded the data in line with these themes. We did not pre-determine any categories other than barriers and facilitators and allowed the themes to emerge from the data itself.

We started the analysis whilst still carrying out our interviews, both so that we could refine the interview questions if necessary and so that we could identify any gaps and attempt to address these through recruiting new participants. The notes from the interviews were divided out among the team who then worked jointly to identify the emerging themes. We then shared our findings with key stakeholders to ensure that they presented an accurate picture of what we had been told.

We developed scenarios and personas based on the themes emerging from the analyses. These are hypothetical situations and characters that depict how things might be in relation to data sharing and linkage in five years time. For each of the main barriers and facilitators we identified, we discussed what the outcomes might be depending on progress in these areas. We then created scenarios based on these possible outcomes and the interplay between them. The scenarios depict a range of outcomes from one where progress has been made in all areas to one where data linkage and sharing has been deprioritised and there has been a lack of any real progress at all.

Annex B: Organisations and teams that contributed to this review

Over the course of the review, the Office for Statistics Regulation gathered information, and received feedback on our analysis and recommendations, from the organisations and teams listed below. In additional to those listed below, we also discussed the report with a small group of academics and independent researchers, each of whom have expertise and interest in the social impacts of technology and who generously gave their time as individuals.

Organisation	Team/Project	Contributed to review	Provided feedback on report
Administrative Data Research UK (ADR UK)		Y	Y
Cabinet Office	Central Digital and Data Office	Y	Y
Data and Analytics Research Environments UK (DARE UK)		Y	Y
Department for Education	National Pupil Database	Y	
Department for Education	Data Improvement Across Government Project	Y	
Department for International Trade		Y	
Department for Levelling Up, Housing and Communities	Better Outcomes Through Linked Data Project	Y	
Department for the Economy Northern Ireland		Y	Y
Department for Work and Pensions		Y	
Department for Science, Innovation and Technology	Centre for Data Ethics and Innovation		Y
Economic and Social Research Council		Y	
Government Office for Science			Y
Information Commissioner's Office			Y
Ministry of Justice	Better Outcomes Through Linked Data Project	Y	
Ministry of Justice	Splink Data Linkage Package	Y	
National Data Guardian			Y

NHS Digital		Y	
Northern Ireland Statistics and Research Agency			Y
Office for National Statistics	Analysis Function		Y
Office for National Statistics	Data Linkage Hub	Y	
Office for National Statistics	Data Science Campus		Y
Office for National Statistics	Integrated Data Programme	Y	Y
Office for National Statistics	National Statistician		Y
Office for National Statistics	Public Policy Analysis	Y	
Office for National Statistics	Methodology and Quality	Y	
Office for Standards in Education, Children's Services and Skills	Grading and Admissions Data for England Data Sharing Project	Y	
Office of Qualifications and Examinations Regulation	Grading and Admissions Data for England Data Sharing Project	Y	
Public Engagement in Data Research Initiative (PEDRI)			Y
Research Data Scotland		Y	Y
SAIL Databank		Y	
SAS UK		Y	
Scottish Government		Y	Y
Thames Valley Together		Y	
UK Longitudinal Linkage Collaboration		Y	
UK Research and Innovation		Y	
UK Statistics Authority	Central Policy Secretariat		Y
UK Statistics Authority	Data Governance, Legislation and Policy		Y
Welsh Government	Administrative Data Research Unit - Wales	Y	Y