Building a Justice Data Infrastructure

Opportunities and Constraints

Stergios Aidinlis, Hannah Smith,
Abi Adams-Prassl and Jeremias Adams-Prassl

September 2020
AI FOR ENGLISH LAW
UKRI GRANT
No ES/ S010424/1

INDUSTRIAL STRATEGY
UK Research and Innovation
UNIVERSITY OF OXFORD

Contents

Click to navigate

Executive Summary

The present report identifies the legal opportunities and constraints that are inherent in creating a robust data infrastructure in the context of the HMCTS digital reform programme. In the advent of increased private-sector involvement in dispute resolution and the prospect of automating judicial decision-making, we explore the potential of justice data to yield substantial benefits for Government and for users of the court system, broadly conceived.

By reviewing the applicable legal framework in the UK, as well as international best practice in creating data infrastructures in the public sector, we elaborate on the required governance arrangements for the promotion of research that will enhance knowledge of the justice system, while at the same time safeguarding the fundamental rights of data subjects, including the most vulnerable.

Our analysis is divided into four distinct, yet overlapping, stages of research-related data processing: collection (I), preparation and linkage (II), access (III) and retention/re-use (IV). We compare contrasting approaches to classifying users of the data infrastructure and advocate for a classification around 'contribution to the public interest'. We elaborate on the implications of this classification in the justice context, considering the types of data that will need to be collected, as well as the potential linkages between HMCTS data and data belonging to other departments. Particular approaches used in international practice, e.g. the use of a trusted-third party (TTP) to minimise identifiability risks when linking datasets, are evaluated and their application to the particular context of a Justice Data Infrastructure considered.

We argue that the interplay between data protection and human rights law provides crucial insights for the responsible research use of justice data: both HMCTS and interested researchers must adhere to key requirements including 'proportionality', 'security', and 'transparency'.

To achieve adequate oversight of all parties' conformity with these principles, we suggest allocating substantive responsibilities to a governance structure. We recommend either the creation of a new body or the investment of further resources into the existing HMCTS Data Access Panel (DAP) in that regard. We discuss the application of the principles by reference to hypothetical case studies of requested data access. A set of concluding recommendations seeks to support the on-going evaluation of HMCTS reform and production of high-quality knowledge about our justice system.

Acknowledgments

We are very grateful to UKRI for generous financial assistance, Professor John Armour, Dr Natalie Byrom, Dr Jassim Happa, Jack Hardinges, Jessica Brown, Mark Sutcliffe, Dr Joe Tomlinson, and Dr Judith Townend for helpful feedback on this report and research strategy. We thank all participants at workshops held in September 2019 and March 2020 for their comments, suggestions, and fruitful discussion.

Introduction

This is a time of monumental change for the UK legal system. In 2016, Her Majesty's Courts and Tribunals (HMCTS) initiated an ambitious programme of court reform, investing £1bn into new technologies to transform the operation of the UK courts and tribunals. Recent reports highlight the wide range of areas affected by the digitalisation agenda: from criminal and civil litigation through to social security tribunals and family law. The full range of legal processes are affected by the shift towards digital justice, from online filing through to fully-fledged online trials and continuous online hearings. Both the scale and speed of change are considerable: official timelines suggest an ambition to resolve most civil disputes through an online court by the early 2020s.[1] The Covid-19 pandemic will only serve further to accelerate the implementation of this agenda.[2] The prospect of complete automation is the next logical step advocated for by some, with AI directly substituting for judicial decision-making.[3] Indeed, private sector companies are increasingly developing technology that offers automated online dispute resolution (often on an opt-in basis).

The shift towards digital justice will create a wealth of new data on legal processes and outcomes. There is potential for this data to yield substantial benefits for Government and for users of the court system. However, many challenging questions remain to be tackled: what data should be captured? How should it be stored? How should it be processed and linked to data from other data owners? Who can access which elements under what governance arrangements?

Such questions have captured the imagination of both policy-makers and academics. They are important both for the efficient operation of the digitalised justice system and for any evaluation of HMCTS reforms. Independent and robust empirical research should assess the impact of the reform on the ability of court users to access a fair justice system. HMCTS aims to facilitate such accountability, publicly committing themselves to streamlining the provision of access to justice data for the purposes of independent research. They have also recently collaborated with The Legal Education Foundation (LEF) in the production of a report that, *inter alia*, identified the major priorities of various stakeholders with regard to the types of data that will need to be collected by HMCTS for evaluation and accountability purposes.[4]

One of the recommendations of the LEF report was that HMCTS should 'dedicate resource to reviewing national and international best practice, existing legal frameworks (…) testing the acceptability of different models with stakeholders and the public'.[5] As part of a UKRI-funded research project, AI for English Law, we aim to contribute to this enquiry by identifying the main constraints and opportunities present in designing a Justice Data Infrastructure from a legal and governance perspective. Analysis is based on our research that involved a desk-based survey of international academic and 'grey' literature, as well as discussions and workshops with representatives from key stakeholders.

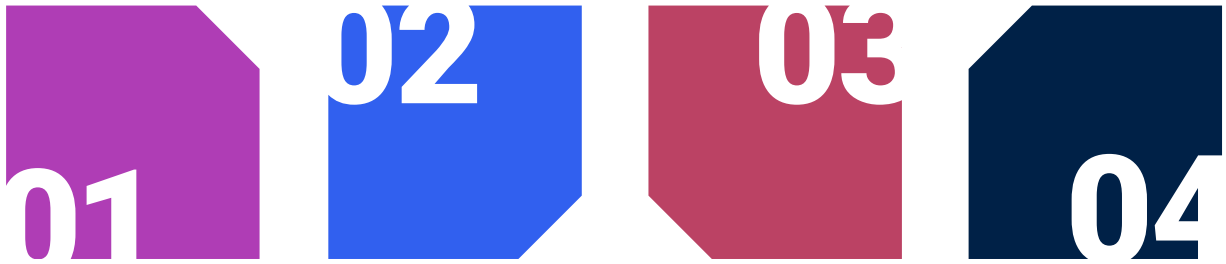
1. HMCTS, 'Reform Update, Summer 2019' https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/806959/HMCTS_Reform_Update_Summer_19.p
2. For a first overview, see N Byrom et al, *The impact of COVID-19 measures on the civil justice system* (Civil Justice Council, London, May 2020
3. For an overview of the arguments and concerns around the automation of justice, see R Binns et al, '*It's Reducing a Human Being to a Percentage: Perceptions of Justice in Algorithmic Decisions*' Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18).
4. N Byrom, 'Digital Justice: HMCTS data strategy and delivering access to justice' (October 2019) https://research.thelegaleducationfoundation.org/wp-content/uploads/2019/09/DigitalJusticeFINAL.pdf
5. *Ibid 6*.

Our findings are relevant to the question of which principles should guide the design of HMCTS' Justice Data Infrastructure in the light of applicable legal frameworks and international best practice on designing data infrastructures for research and evaluation purposes. In this report, we present an overview of the legal and governance requirements for a data infrastructure that will collect, curate, and retain justice data with a view to stimulating independent research on the performance of the Courts and Tribunals system by reference to the core aims of the HMCTS reforms.

To clearly establish the connection between these aims and user access to justice data, we advocate for a classification of data users around the notion of 'public interest'. We argue that the interplay between data protection and human rights law is favourable for on-going research use of justice data that furthers the 'public interest', provided that both data custodians and researchers adhere to key requirements including 'proportionality', 'security' and 'transparency'.

We elaborate upon the ways in which legal and ethical principles need to be incorporated into any data infrastructure by-design, as well as how they need to be communicated to researchers and the general public. Several approaches used in international practice, e.g. the use of a trusted-third party (TTP) to minimise identifiability risks when linking datasets, are evaluated and their application to the particular context of justice data is considered. Finally, yet importantly, we stress the significance of allocating substantive responsibilities to a governance structure that will oversee the application of the presently discussed principles. We recommend either the creation of a new body or the investment of further resources into the existing HMCTS Data Access Panel (DAP) in that regard.

The structure of this report reflects four distinct, yet overlapping, stages of HMCTS research-related data processing:



01

Data Collection.

In this section:

1. Defining and Classifying Users
2. Applying the Classification around ‘Public Interest’ in Practice
3. Designing Data Collection Principles

Data Collection

The present report discusses the legal opportunities and constraints in designing a data infrastructure that will support justice data sharing for research and evaluation purposes in the context of the HMCTS digital reform in the United Kingdom.[6] In doing so, it draws on the applicable UK legal framework for data sharing, as well as on international best practice in establishing and maintaining public-sector, administrative data sharing governance arrangements. In this first chapter, we make the case for defining and classifying users, also discussing the potential application of our proposed solution to a number of illustrative case studies. We then proceed with the legal, ethical, and technical challenges during the first stage of data processing, i.e. *collection*.

1. Defining and Classifying Users

HMCTS have committed to facilitating 'independent research' on the reform programme. How, however, are we to distinguish 'independent' from 'non-independent' research? Are all types of independent research legitimate and acceptable? In classifying different categories of data users, many existing initiatives distinguish between academic and commercial researchers.[7] For example, data-intensive research investments such as the UK Economic and Social Research Council-funded Administrative Data Research Network (ADRN) used this delineation to confine its data sharing to academic researchers. The ADRN justified this approach, in part, by reference to public attitudes.[8]

KEY PIECES OF LEGISLATION FOR THE DATA COLLECTION STAGE

- *EU General Data Protection Regulation 2016/679 (GDPR)*
  - » *Articles 5(1)(b), 6(1)(a), (c) and (e), 6(3)(b), 6(4)(c) and 89*
  - » *Recitals 43 and 50*
- *Digital Economy Act (DEA) 2017 UK*
  - » *Sections 64-70*

6 'Research' and 'evaluation' are used in this context in a broad manner, covering, in principle, both academic and non-academic, as well as both internal and external uses of government data in the interest of improving social-scientific knowledge about the operation of the justice system through data-intensive analytical techniques. This is consistent with the broad definition of 'research' in recital 159 GDPR: 'For the purposes of this Regulation, the processing of personal data for scientific research purposes should be interpreted in a broad manner including (…) privately funded research'.

7. S Price, '*Academic and commercial research: Bridging the gap*' (2015) 12(2) Participations 168.

8. IPSOS Mori, '*Commercial Access to Health Data*' (2016) https://www.ipsos.com/ipsos-mori/en-uk/commercial-access-health-data?search=commercial%20access%20health

The following box presents the employment of this type of definition by the ADRN:

ACADEMIC VS COMMERCIAL RESEARCH

The Administrative Data Research Network (ADRN) was funded by the Economic and Social Research Council (ESRC) from 2013 to 2018. It brought together experts from all over the UK to organise an infrastructure that allows social researchers to use administrative data in a safe setting, with the proper security measures in place.9

In 2015, the Northern Irish Centre of the ADRN commissioned a survey to investigate public attitudes in Northern Ireland towards data sharing for both research and care purposes.10 The survey results demonstrated that individuals drew distinctions based on the type of organisation, with higher numbers trusting academic researchers (72%) over trusts and charities (51%) or commercial and insurance companies (41%) to use their data appropriately. Furthermore, one in two participants (50%) expressed a preference for more robust safeguards to be in place where their data was processed by commercial companies as opposed to academic researchers.

These findings suggest that an entity's organisational identity as an academic or a commercial research entity may be a crucial factor to consider when granting access to any data infrastructure. Even when commercial access to data is not precluded, many existing data infrastructures still utilise this distinction to impose different requirements upon academic researchers and commercial organisations. One example is the UK Data Service and its Secure Lab, which employs different pathways in granting access to its data based on whether the user is a UK academic researcher, a local government authority, a charity, a non-UK user or a commercial organisation.[11]

The HMRC Data Access Panel also predominantly confines access to its data to non-commercial entities.[12] Originally, the HMRC panel accepted applications only from researchers based at a UK academic institution or government department. In April 2018, however, it opened up access to its data to commercial research groups, although such data sharing is restricted to projects commissioned by a government department.[13]

9.. ESRC, 'Administrative Data Research Network' (2018) adruk.org

10. G Robinson et al. '*Public Attitudes to Data Sharing in Northern Ireland: Findings from the 2015 Northern Ireland Life and Times survey*' (2018) pure.ulster.ac.uk/en/publications/public-attitudes-to-data-sharing-in-northern-ireland-findings-fro

11. UK Data Service, 'Registration' ukdataservice.ac.uk/get-data/how-to-access/registration/commercialusers.aspx

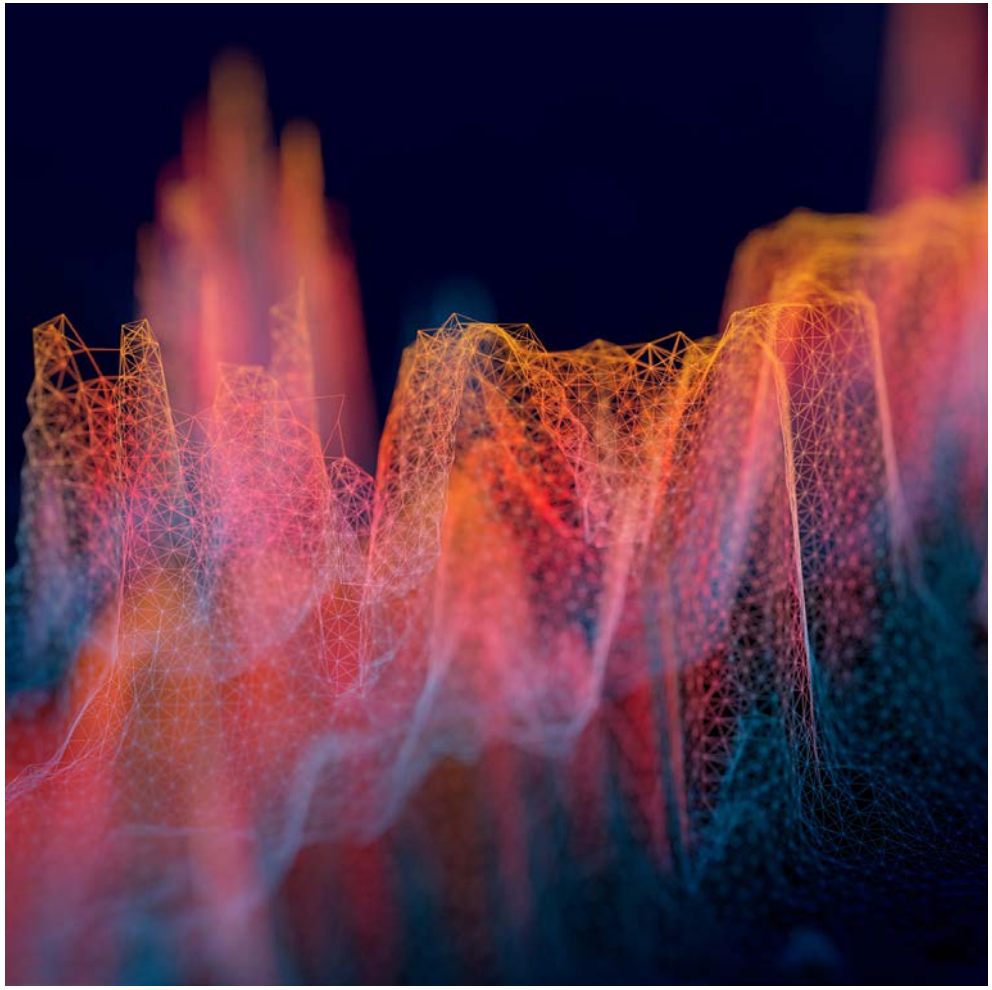
12. R Welpton and M Wright, '*HMRC Datalab: Engaging with the External Research Community*' (2012) data-archive.ac.uk/media/425160/hmrcdatalab.pdf

13. HMRC, 'Research at HMRC' gov.uk/government/organisations/hm-revenue-customs/about/research

There are also cases of data infrastructures which are designed to accommodate the needs of a specific category of user, who may not be an academic researcher. One example of this approach is the Justice Data Lab at the Ministry of Justice. It facilitates access to re-offending data with the aim of enabling users better to assess various intervention approaches and efforts to rehabilitate offenders.[14] In consequence, the majority of its users have been voluntary and community organisations, social enterprises, and private businesses given their key role in creating, funding, and implementing these initiatives.

Whilst organisational identity is a well-understood and widely-used criterion, it also suffers from a number of weaknesses, particularly when considered in light of modern data processing activities. As a criterion, it is insufficiently nuanced to reflect the intricacies of multi-stakeholder data sharing for research. A series of recent data governance controversies demonstrate that the organisational identity of a user is insufficient alone to justify different treatment towards the sharing of data.[15] Moreover, a rigid delineation of users based on organisational identity might pre-empt the potential of certain types of organisations to contribute to the production of high-quality research that may be used to improve public service delivery.[16]

With these concerns in mind, one alternative to organisational identity is to classify users based on their *intended use of data*, i.e. their purpose behind accessing the infrastructure's resources and performing data analysis on them. The UK Data Service has followed this approach in articulating different requirements for the users of its Secure Lab distinguishing between *commercial* and *non-commercial use*.



14. F Lyon et al, '*Opening access to administrative data for evaluating public services: The case of the Justice Data Lab*' (2015) 21(2) Evaluation 232.
15. P Lewis et al, 'Cambridge Analytica academic's work upset university colleagues' (2018) theguardian.com/education/2018/mar/24/cambridge-analytica-academics-work-upset-university-colleagues
16. E.g. on the data sharing agreement between Amazon and the NHS, see Department of Health and Social Care, 16 October 2019, 'Amazon Master Content License Agreement' www.contractsfinder.service.gov.uk/Notice/919533b2-4d46-4c72-bf2b-4e320cff572e cf. Ada Lovelace Institute, 'Health Data Partnerships: Amazon/Department of Health and Social Care – Ada's view' (16 December 2019) adalovelaceinstitute.org/health-data-partnerships-adas-view

COMMERCIAL USE VS NON-COMMERCIAL USE

In principle, commercial organisations can access and download or order data from the Secure Lab of the UK Data Service.[17] *The conditions under this can happen, however, are determined by their intended use of the data.*

A distinction is drawn between commercial and non-commercial use. The former corresponds to research where the 'direct objective is to generate revenue and/or where data are requested for sale, resale, loan, transfer or hire' (emphasis added).[18] *The latter refers to data uses that seek to support 'a public good (…) i.e. an activity which widens access to information sourced from our collection and has social or economic benefit' as a result of the use.*[19]

When the intended use of its data is deemed to be commercial, the UK Data Service must seek the permission of the 'data owner' and the data infrastructure user must agree to a commercial agreement between it and the UK Data Service.[20] *Commercial users are also prevented from accessing some of the datasets and their requests are subject to an administrative charge.*

While data use is a more dynamic and context-specific definitional approach, there is still the risk of ambiguity when trying to pin down the meaning of subjective concepts such as the 'intended use of data'. Using the UK Data Service example, it is not entirely clear how a 'direct' objective to generate revenues may be distinguished from a relevant 'indirect' objective. There are certain organisational requirements and limitations associated with the outcomes of interpreting such an objective and they can be onerous for potential users.

Although interpreting and defining general categories is part-and-parcel of a data infrastructure's day-to-day governance, more clarity on how the users are defined would be beneficial when creating a new infrastructure. This is all the more crucial in the light of empirical findings on the existence of a 'culture of caution' in the UK public sector,[21] with many data owners reluctant to disclose administrative datasets for research and evaluation purposes. Ambiguity may make a more cautious approach preferable for data owners, whilst potentially hampering beneficial for the public data-intensive research.

17. UK Data Service (n 11).
18. *Ibid*.
19. *Ibid*.
20. 'Data owner' refers here to the public body responsible for managing the relevant dataset since ownership or property is not a legal concept in the case of administrative data. On this, see Your Response Ltd v Datateam Business Media Ltd (2014) EWCA Civ 281 (Moore-Bick LJ) [29-34].
21. G Laurie and L Stevens, '*Developing a Public Interest Mandate for the Governance and Use of Administrative Data in the United Kingdom*' (2016) 43 JLS 360, 362; Richard Thomas and Mark Walport, `*Data Sharing Review Report*' (2008) amberhawk.typepad.com/files/thomas-walport-datasharingreview2008.pdf

For these reasons, ***we propose a classification that utilises a project-based approach***, considering the project's substantive relevance to the public interest. A similar approach is followed by the Office for National Statistics (ONS) in the UK, where the relevant delineation has strict repercussions for data access to its Virtual Microdata Laboratory (VML).[22] As discussed in the next sub-section, this definition utilises elements of the previous approaches, yet seeks to provide a clearer overarching framework in assessing the acceptability of data access requests. We propose the following classification of data users:

A CLASSIFICATION AROUND THE 'PUBLIC INTEREST'

We propose the following classification of potential users:

A. Public interest research – when the research project exclusively serves the public interest.

B. Proprietary research – when the research project exclusively serves private interests of a proprietary nature.

C. Hybrid research – when the research project serves ***both*** *the public interest and private interests of a proprietary nature.*

By law, HMCTS are currently allowed to grant access to their data for research purposes only when such research is 'in the public interest'.[23] This clearly includes research classified as 'public interest', commonly associated with academics, but not 'proprietary' research, commonly associated with corporate entities. However, the notion of 'public interest' has not yet been authoritatively interpreted by the Courts and thus there is some ambiguity in the treatment of 'hybrid' research. These challenges are inherent in the interpretation and reconciliation of normative categories that are not necessarily mutually exclusive, yet often create conflicts in practice.[24] Since the notion of 'public interest' is notoriously elusive, prone to various definitions based on the context in which it is being utilised,[25] and subject to different interpretations by individuals,[26] it is important to consider how the proposed classification would work in practice.

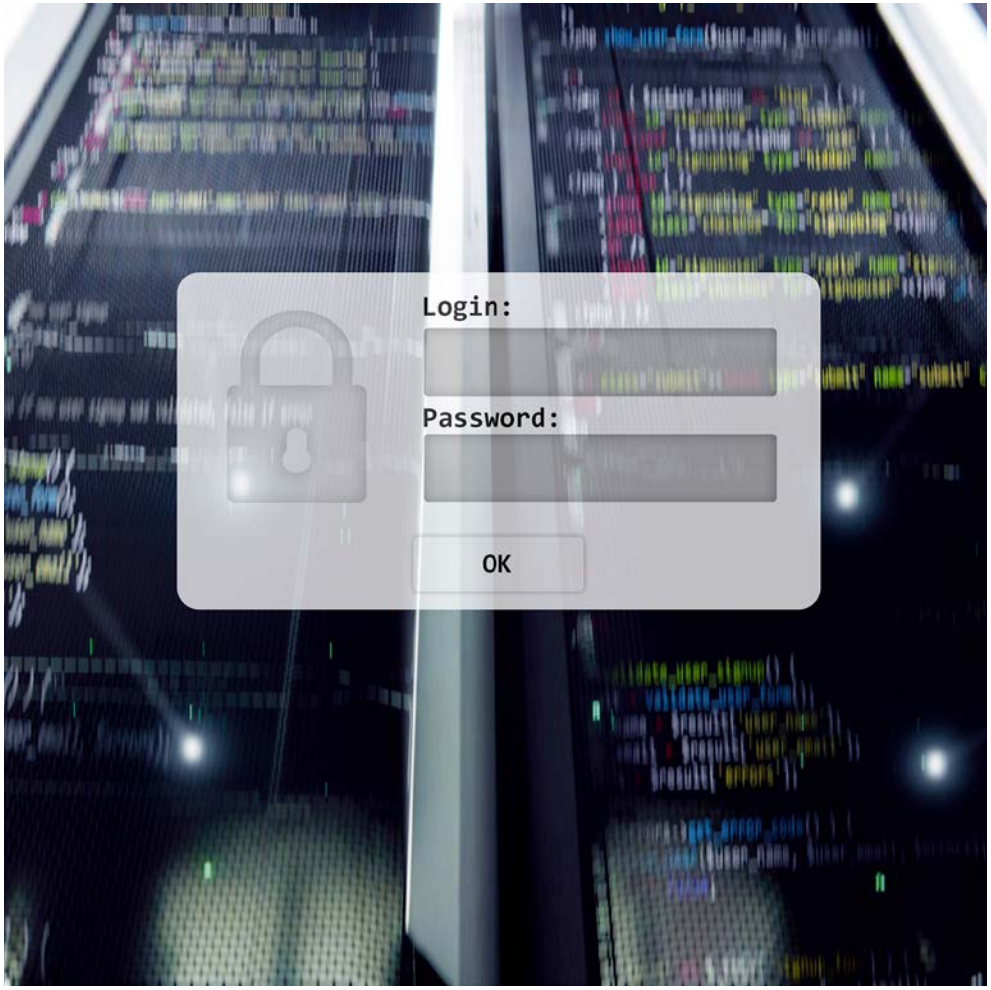
22. '*Access is only granted for research that serves the public good*', ONS, 'Approved Researcher Scheme' ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme
23. Under ss. 64, 71(4) Digital Economy Act 2017.
24. G Laurie et al, '*Charting Regulatory Stewardship in Health Research*' (2018) 27(2) Cambridge Quarterly of Healthcare Ethics' 333, 340.
25. Even data infrastructures do not share a common definition of the 'public interest', with the UK Data Service giving a broad definition of 'widening access to knowledge and benefiting the economy and the society', whereas the ONS gives a much more specific 7-point list of the forms that the public interest can take, *supra* (ns 11 and 21).
26. P Carter et al, '*The social licence for research: why care.data ran into trouble*' (2015) 41 Journal of Medical Ethics 404.

2. Applying the Classification around 'Public Interest' in Practice

The intricacies of developing and consistently applying a definition of the public interest necessitate a responsible governance body or Data Access Panel to have oversight of the proposed classification.[27] Such a panel would oversee the drafting of data sharing agreements, monitor whether users are compliant with the terms, and enforce these agreements, to ensure that the public interest is served.[28] To provide a blueprint for this task, we articulate key factors to be considered in the implementation of the proposed classification and discuss their application by reference to a number of hypothetical case studies.

A few caveats need to be mentioned at the outset of this discussion. First, this is an illustrative rather than an exhaustive determination of the factors that need to be considered, since the factors related to the acceptability of an access request are numerous and may be contingent on the context. Second, the following discussion is not intended to give a final determination of whether we believe that such uses of data should be permitted or denied. Instead, it serves to elucidate some of the relevant factors, as well as highlighting the breadth of expertise we believe is necessary within the Data Access Panel to ensure an appropriate evaluation of the ethical, legal, and societal implications of a data access request.

Based on the above, we suggest that the following factors need to be considered: (a) the ***identity of the actors*** seeking access to the data, (b) the ***types of data*** being requested, (c) the ***intended purpose*** behind accessing the data, and (d) the possibility of subsequent ***re-use and re-processing***. These factors are outlined in more detail in the following table.



27. We elaborate on this requirement on the third chapter of the present report, *infra*, III.1.
28. With Section 170 of the Data Protection Act 2018 (UK) enshrining a criminal offence of knowingly or recklessly obtaining, disclosing, procuring or retaining personal data without the consent of the data controller, and the sale or offering for sale of that data.

KEY QUESTIONS IN APPLYING THE 'PUBLIC INTEREST' CLASSIFICATION

1. The identity of the actors seeking access to the data

a. How should the institutional identity (e.g. charitable status, academic status, or commercial affiliation) of an applicant influence the panel's decision?

b. How should the accreditation of researchers influence the Panel's decision? What types of expertise (e.g. qualifications, publications, work experience) should be relevant to the Panel's decision?

c. Should the independence of a researcher be of relevance to a decision? If so, how could this independence be established and measured?

2. The types of data requested

a. Is the requested data necessary for the proposed project? Is it sufficient? Could a more or less granular approach be implemented? Should it?

b. To what extent does the extraction of the requested data create novel risks for the interests of individuals, including, but not just limited to, their privacy?

c. Does the request respect the principles of data minimisation and security?

3. The intended purpose

a. Does the intended purpose accord with the definition of the public interest in law (e.g. s 7 of the SRSA 2007, s 71(4) of the DEA 2017 read together with the UKSA code of practice)?

b. Is the intended purpose consistent with data subjects' reasonable expectations as to how HMCTS could use their data?

c. Have the risks associated with the request been evaluated in a comprehensive manner? Do these outweigh the potential benefits? Are the risks and benefits equal for all social groups?

4. The possibility of subsequent re-use or re-processing

a. Does the request seek to retain the requested data in the interest of future re-use and re-processing?

b. Does the request include specific plans for the re-use of the data e.g. in future research addressing particular questions?

How would these factors apply in practice?

To explore this question, we will analyse international best practice in implementing a lawful and ethical data sharing policy with regard to these factors throughout this report. We furthermore include a number of hypothetical case studies to illustrate the application of the outlined factors.

CASE STUDY 1: CHARITY ACCESS

A charity, working in the area of violent offenders' rehabilitation, approach HMCTS and request access to the Justice Data Infrastructure. They plan to fund an independent, accredited researcher to pursue a research project on the connection between reoffending and unemployment, having already secured a permission from the DWP to use employment data for this purpose. They plan to publish this research in a report on their website, as well as communicate it to the press, to raise awareness about the support needs of some of the most vulnerable users of the justice system.

A Data Access Panel would consider a number of key issues in response to this request. First, it would seek to establish that the disclosure is in the public interest in respect of the identity of the actor submitting the request. In this case, exploring the connection between re-offending and unemployment from the perspective of a charity that will then publicise the findings bears potential to 'extend understanding of social (…) trends (…) by improving knowledge'.[29]

Second, with regard to the types of data being requested, the panel would seek to minimise identification risks when linking HMCTS with DWP data by applying, potentially with the help of a Trusted Third Party, pseudonymisation procedures to separate content data (e.g. conviction judgment or employment contract) from identifiable demographics (names, addresses etc). Holding that this is consistent with the charity's analytical aims would be of utmost importance. Third, the panel would seek to clarify whether the charity plans to retain datasets for future use by them or the HMCTS. In this case, the panel would need to be convinced that the charity only holds non-identifiable data that is necessary for its aims, or that very strict data security controls are applied to the handling of any identifiable information. Chapters 2, 3 and 4 of the report, on data linkage, access, and retention, elaborate on how these requirements, which stem from the Digital Economy Act 2017, the EU GDPR, and human rights law, have been met in international best practice.[30]

29. UK Data Service (n 11).
30. *Infra*, II, III and IV.

CASE STUDY 2: LEGAL START-UP ACCESS

A start-up, currently operating as a not-for-profit social enterprise, approach HMCTS and request access to the Justice Data Infrastructure. They are interested in developing novel technology solutions to improve access to the reformed justice system by digitally-excluded litigants. Their mission is to alleviate power imbalances that can occur when vulnerable users face well-resourced respondents who can afford expensive legal representation. In the long run, the start-up would consider the marketisation of their products, with governments, international organisations and non-Government organisations amongst their potential clients.

From an organisational identity perspective, a key question in this case study is whether the potential marketisation of the start-up's products in the future creates complications for data access. A data access panel would have to elaborate on what HMCTS interprets as a 'contribution to the public interest'. The start-up's activity aims to contribute towards creating an evidence-base for 'decisions which are likely to significantly benefit the (…) quality of life of people in the UK', particularly the individuals who might face difficulty accessing the digital justice system. Regarding the intended purpose of data use, the data access panel would discuss whether and how such a use contributes to the HMCTS reform mission to create a more efficient and accessible justice system.

31. *Infra*, III
32. *Infra*, III and IV.

Chapter 3 of the report elaborates on how the 'public interest' has been operationalized in international best practice.[31] Having established that access can be granted, it would be for the panel to require that the start-up liaise with the UKSA to ensure all the relevant individuals seek and attain the necessary accreditation, as well as ethical approval from an appropriate body (either the NSDEC or the HMCTS panel itself). In respect of data re-use and re-processing, the panel would also seek formally to agree the status of any retained data with the start-up, clarifying whether the start-up will be allowed to hold any datasets after deciding to commercialise their product, and the anonymisation techniques they should apply before doing so. Chapters 3 and 4 of the report elaborate on how these requirements, which stem from the Digital Economy Act 2017, the EU GDPR, and human rights law, have been met in international best practice. [32]

CASE STUDY 3: LAW FIRM

A boutique law firm, specialising in the application of artificial intelligence to legal practice, approach HMCTS and request access to the Justice Data Infrastructure. They are primarily interested in systematising risk factors in litigation and developing an algorithmic model that can then be used either to provide a competitive advantage against other law firms, or be patented and sold to large, multinational law-firms as a software package. They believe that they can achieve the development of such a product through mining HMCTS databases.

This scenario raises several challenges for the HMCTS data access panel. First, from the perspective of the requesting actor's identity, establishing the contribution of this proposed disclosure to the public interest will be more difficult, considering the proprietary aims of the law firm. With regard to the intended purpose of data use, the data access panel would seek to understand how the law firm's proposed research in systematising risk factors in litigation could be beneficial for HMCTS's organisational mission and to what extent the law firm would be willing and able to facilitate such benefits. If the panel were satisfied, they could potentially provisionally grant access. Second, in respect of the requested data, considering the interest of the law firm in systemic qualities of the justice system, the panel would seek to minimise the amount of identifiable data that the law firm could access, applying the appropriate pseudonymisation techniques. Before any access could be granted, the panel would seek to establish that the law firm's researchers are accredited and reliable, as well as that the ethical implications of the disclosure have been considered by an appropriate body. Finally, the law firm would have to formally agree that no identifiable data would be retained within their produced software package. Chapters 3 and 4 of the report elaborate on how these requirements, which stem from the Digital Economy Act 2017, the EU GDPR, and human rights law, have been fulfilled in international best practice.[33]

Having elaborated on these hypothetical case studies, it is worth emphasising again that implementing a consistent data sharing policy will take time and potentially require a lot of engagement between the HMCTS data access panel and requesting researchers. Further, bearing in mind the often exploratory character of data research, it is important that the adopted definition of the public interest does not preclude research that may bear significant, even if yet undefined, potential benefits to the public interest by identifying unknown patterns, links, and correlations that could support a more efficient delivery of justice.[34] The following chapters will offer helpful examples of international best practice, within which legal and ethical requirements have been put in place as guiding principles of relevant research data infrastructures.[35] We now turn to the first stage of data processing, i.e. *data collection*.

33. *Infra*, III and IV.

34. X Jin et al, '*Significance and Challenges of Big Data Research*' (2015) 13 Big Data Research 5

35. A helpful example in that regard would be the Health Research Authority's (HRA) legal and ethical review of projects, involving such bodies as the Confidentiality Advisory Group (CAG) and Ethics Committees, see HRA, 'What approvals and decisions do I need?' hra.nhs.uk/approvals-amendments/what-approvals-do-i-need

3. Designing Data Collection Principles

The first stage of data processing in a data infrastructure involves *collecting* or *re-purposing data* that is already being collected for research and evaluation purposes. In this report, we will be assuming that the data infrastructure will be collecting or processing personal,[36] individual-level data in the interest of achieving new insights that will be used to improve knowledge about the justice system.[37] This is not to say, however, that data used for research will be identifiable at *the data user* level, i.e. when individual researchers access and analyse them.[38] Following chapters elaborate on these issues. The present chapter focuses on challenges around mapping collected data for research purposes, as well as adhering to requirements of lawful and fair processing, purpose limitation, and data minimisation.

A. Data Mapping

Data mapping refers to a scoping exercise about the kind of information that is already collected or will need to be collected for the first time within a Justice Data Infrastructure. The idea of 'mapping' has been used to capture not only the data assets in the strict sense, i.e. the information, but also the 'ecosystem of organisations' that will be collaborating to ensure that the data infrastructure produces maximum value for its users.[39]

While 'data mapping' will be an ongoing task for HMCTS, we elaborate here on the intended output of this exercise, as well as the key types of information that mapping should begin from and the importance of consulting the relevant stakeholders.

The HMCTS data infrastructure should gradually build on the construction of a *data catalogue* or *prospectus*, in accordance with best practice in the UK with regard to identifying target datasets for research use.[40] The UK Data Service has created a relevant catalogue of their data, containing 7483 studies and 73 series, of which 165 are classed as 'controlled' (personal data) and require approval by the data owners before they can be accessed.[41]

Similarly, the HMRC DataLab regularly updates a list of datasets available to researchers.[42] Like the HMRC DataLab, the SAIL databank also publishes a list of its available datasets, distinguishing between 'core datasets' and 'core restricted datasets', with the latter requiring an additional permission from their data owners to be used for research.[43] It is helpful to consider how other initiatives and data centres have conducted data mapping exercises. The Justice Data Lab[44] at the Ministry of Justice is an example where data mapping has led to the concrete identification of target datasets.

36. On personal data 'concerning an identified or identifiable natural person', by contrast to anonymous information which 'does not relate to an identified or identifiable natural person', see Recital 26 Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/ (General Data Protection Regulation) [2016] OJ L119/1, which will be cited hereafter as 'the GDPR'.
37. Whilst a focal issue for discussions of justice data governance, the treatment of judicial judgments is not considered in this report due to the particularities of their production and dissemination, as well as ownership and management as justice data. For a treatment of Judgments and tribunal decisions as justice data, see the ongoing research by led by Dr Judith Townend, '*Justice System Data research: a comparative study*' sussex.ac.uk/law/research/projects/justicesystemdataresearchacomparativestudy
38. On 'identifiability' in data protection law as a context-specific concept and the so-called 'functional anonymisation' of data when researchers access them, see M Mourby et al, '*Are "pseudonymised" data always personal data? Implications of the GDPR for administrative data research in the UK*' (2018) 34(2) Computer Law & Security Review 222; M Elliot et al, The Anonymisation Decision-Making Framework (UKAN, 2016).
39. Open Data Institute, 'Mapping Data Ecosystems' theodi.org/article/mapping-data-ecosystems 4.
40. Creating an 'external-facing data catalogue' was also one of the recommendations of the TLEF report, Byrom (n 4) 7.
41. UK Data Service, 'About our data' ukdataservice.ac.uk/get-data/about.aspx
42. HMRC, 'HMRC Datalab datasets available' (updated 17 April 2019) gov.uk/guidance/hmrc-datalab-datasets-available
43. SAIL Databank, 'SAIL datasets' saildatabank.com/saildata/sail-datasets
44. Lyon et al (n 14).

MAPPING KEY DATASETS AT THE JUSTICE DATA LAB

Aiming to enable better access to reoffending data for providers of offender interventions so that the latter can assess their impact on reoffending behaviour, the Justice Data Lab (JDL) set out to clarify the key information to be collected and linked.[45]

The first consideration was the type of data that is to be collected, linked, and shared with data infrastructure users. The JDL collects specific variables from providers of offender interventions: 'first name, surname, date of birth, gender and dates that refer to the participation in the intervention or the sentence that led to this participation'.[46]

This data is then linked to Ministry of Justice administrative datasets such as the Police National Computer, reoffending databases and Offender Assessment Information.[47] *A comparison group is created to 'match to and analyse aggregate reoffending information'. Crucially, the three internal Ministry of Justice datasets are either owned or held by Justice Statistics for research and analytical purposes. Thus, no new data needed to be sourced for the data infrastructure operation, with the relevant data providers being aware and approving of this use. This was also crucial in respect of resource allocation, both in terms of staff and money.*

Data that is already captured as part of routine case management, e.g. key identifiable variables of litigants and necessary case management information, would clearly fall within the desired scope. The collection of data on the outcome of a claim would be of similar importance. For example, recording, where applicable, the stage of withdrawal could highlight what explains delays in court proceedings and facilitate the design of appropriate reforms that will allow timely dispute resolution. Demographic and equalities data must also be collected in the interest of providing targeted support to justice system users with different needs.[48]

Data not strictly necessary for case management, but with a potential to be meaningfully linked across cases, jurisdictions, and departments, is another key consideration. Such data, once linked, could assist in areas such as enforcement of judicial decisions or the prevention of disputes from arising in the first instance. For example, the NHS Wales Informatics Service uses deterministic and probabilistic linkage routines sequentially to develop a matching algorithm to preserve record integrity and identity between Welsh population demographic databases and the anonymised datasets in the SAIL databank.[49]

45. Ministry of Justice, '*Justice Data Lab Data Protection Impact Assessment*' (May 2018) assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709978/jdl-data-protection-impact-assessment.pdf
46. *Ibid* 3
47. *Ibid*
48. *Ibid*
49. K Jones et al, '*10 Years of Spearheading Data Privacy and Research Utility*' dx.doi.org/10.23889/sail-databank.1001101

HMCTS should establish a standing network of stakeholders that will need to be consulted in respect of data that will be collected throughout and after the reform process. A helpful groundwork can be found in a recent report by The Legal Education Foundation (LEF).[50] The LEF consulted various stakeholders from the judiciary, the civil service, academia, and the voluntary / community sector to identify their respective primary needs in terms of HMCTS data collection.[51] The report proposed focusing the data infrastructure's collection strategy on thirteen data points that are related to user vulnerability, e.g. age, disability, gender reassignment or fear of distress connected with the case.[52] This could facilitate the alignment of the collection of justice system data and 'existing legal duties relating to access to and the fairness of the justice system, as well as obligations under the Public Sector Equality duty'.[53]

Identifying the types of information collected by a Justice Data Infrastructure will be an iterative process, in the interest of producing a robust, evidence-based body of knowledge about the everyday workings of the courts and tribunals. There are, however, certain legal requirements that must be met in creating such a systematic and dynamic body of knowledge.

B. Lawful and Fair Processing

Beyond data mapping, it is important to demonstrate the 'fairness' and 'lawfulness' of either collecting data anew, or repurposing already collected data, for research and evaluation use.

The requirements discussed here are designed to ensure compliance with data protection law, and also to safeguard adequate respect for the reasonable expectations of data subjects for privacy and their capacity to exercise their individual rights. In legal terms, the so-called first and second data protection principles, i.e. 'lawful basis' and 'purpose limitation', are relevant here.[54]

More specifically, Article 6 GDPR requires that any data processing activity have a lawful basis, providing an exhaustive list of potential legal bases. For present purposes, the relevant bases are consent [6(1)(a)], necessary processing for compliance with a legal obligation [6(1)(c)], and necessary processing for the 'performance of a task carried out in the public interest or in the exercise of official authority vested in the controller' [6(1)(e)]. Among the three, the 'public interest' / 'official authority' is the preferred ground. Recital 43 of the GDPR cautions against public authorities' excessive reliance on consent for processing due to the imbalance of power between the controller and the data subject. While both the Information Commissioner's Office[55] and the Article 29 Working Party[56] clarify that the use of consent in public sector data processing is not totally excluded by the GDPR, there is an increased burden of proof for the public authority to demonstrate that consent was freely given. This makes consent a less appealing basis in this context, both from a logistical and a legal perspective. In a similar vein, it is not clear that an existing legal obligation *mandates* HMCTS to collect data for research purposes.[57]

50. Byrom (n 4)
51. *Ibid* 25-26.
52. *Ibid* 5.
53. *Ibid* 2.
54. Article 6 (1) and (4) of the GDPR; sections 8, 19 UK Data Protection Act 2018.
55. ICO, 'Consent' ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent
56. A29 WP, 'Guidelines on consent under Regulation 2016/679' (28 November 2017) ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051
57. Even in the case of the Public Sector Equality Duty, the core principles established in *R (Brown) v Secretary of State for Work and Pensions* [2008] EWHC 3158 operate at a higher level.

Hence, invoking the 'public interest' / 'official authority' ground under article 6(1)(e) GDPR is the most appropriate basis for collecting and processing data for research purposes. To ensure the lawfulness and fairness of processing, as well as transparency, there are two clarifications that must be made when relying upon article 6(1)(e) GDPR. First, the data controller must clarify the basis for the processing in domestic law, in accordance with article 6(3)(b) GDPR. Crucially, this domestic legal basis does not necessarily need to be either a statutory,[58] or a legal obligation.[59] The Information Commissioner's Office have clarified that article 6(1)(e) GDPR includes discretionary legal powers.[60]

When considering the aims of the proposed data infrastructure, this requirement seems straightforward. Section 8(a) of the UK Data Protection Act 2018 stresses explicitly that the 'administration of justice' falls within the 'public interest' ground; HMCTS could potentially also rely on the general permissive gateway in sections 64-70 of the UK Digital Economy Act 2017.[61] Second, with a view to demonstrating targeted and proportionate processing in later processing stages, it is important to substantiate a 'public interest' mandate for the data infrastructure's collection strategy from the outset. HMCTS can build on these formulations and, arguably, improve them by drawing on the literature which concerns the potential benefits of administrative data research for public policy-making.[62]

CREATING A 'PUBLIC INTEREST' MANDATE FOR A JUSTICE DATA INFRASTRUCTURE

*The creation of a 'public interest' mandate in practice would entail a policy that convincingly articulates **why** collecting and processing data for research bears great potential to transform the administration of justice and create a more efficient and accessible system. This would be particularly significant to ensure fairness of processing and shape data subjects' reasonable expectations of privacy.*

*Relevant formulations can be observed in the practice of existing data infrastructures. The HMRC datalab, using section 17 of the CRCA 2005 to share data for research, clarifies that data collected for tax purposes can be used for research to 'improve and develop other HMRC services, without seeking the permission of individual taxpayers'.[63] The Justice Data Lab utilises section 14 Offenders Management Act 2007, permitting disclosure of information for the purposes of the 'management of offenders', to justify its data sharing activities.[64] The JDL suggest that such data sharing is **necessary** for the Ministry of Justice to acquire an evidence basis that will guide policy-making in the future.*

58. Recital 41 GDPR: '(…) this does not necessarily require a legislative act adopted by a parliament (…)'.
59. ICO, 'Public task' ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/public-task
60. *Ibid*: 'a public body's tasks, functions, duties or powers'.
61. Providing both the power for public authorities to disclose data for research purposes, and the parameters within which it should be exercised.
62. E.g. Lisbeth Rivas and Joe Crowley, 'Using Administrative Data to Enhance Policymaking in Developing Countries: Tax Data and the National Accounts' (IMF working paper) imf.org/en/Publications/WP/Issues/2018/08/02/Using-Administrative-Data-to-Enhance-Policymaking-in-Developing-Countries-Tax-Data-and-the-46054 Laurie and Stevens (n 16).
63. HMRC (n 13).
64. Lyon et al (n 14).

C. Purpose limitation and data minimisation

It is crucial to consider the data infrastructure's conformity with two further principles: purpose limitation and data minimisation. Purpose limitation comes into play when repurposing already collected data for research and evaluation. The principle of data minimisation can apply both when repurposing data and when collecting data for the first time. With purpose limitation, a tension seems to be created between Recital 50, article 5(1)(b) and article 6(4)(c) GDPR. Recital 50 *inter alia* mentions that:

> 'The processing of personal data for purposes other than those for which the personal data were initially collected should be allowed only where the processing is compatible with the purposes for which the personal data were initially collected (…) Further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be considered to be compatible lawful processing operations'.

Article 5(1)(b) GDPR qualifies this presumption of compatibility by stressing that such further processing 'shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes.' This would seem to guarantee compatibility when e.g. the initial purpose is routine case management and the further processing takes place for research purposes, provided that the safeguards of Article 89 (including data minimisation) are met. Article 6(4)(c) GDPR, however, mandates that data controllers take into account 'the nature of the personal data, in particular (…) personal data related to criminal convictions and offences' when re-purposing data collected for another purpose. It is clear the present data infrastructure may aspire to process this type of personal data for research and evaluation purposes.

The outcome of this interplay is unclear. The former provisions establish a presumption of compatibility, whereas the latter only stipulates one relevant consideration in a broader assessment. Considering that the European Data Protection Board have recently found the conditions for applying this presumption of compatibility 'horizontal and complex',[65] it is safe to assume that any data infrastructure will have to produce a tailor-made policy on the matter. For example, the Justice Data Lab must clarify the precise variables they use to identify sentences and follow-up periods in measuring re-offending, explaining why these are necessary to ensure quality of the analysis.[66] Furthermore, the application of data minimisation safeguards under article 89 GDPR, including de-identification of data when provided to the researchers, will be crucial in demonstrating conformity with the 'purpose limitation' principle. The application of these principles will also be discussed in the last chapter on data retention and re-use.[67]

Finally, yet importantly, a potential challenge that is relevant to data minimisation stems from the Government Digital Service Standard (GDSS). While not a statutory legal requirement, the GDSS is a Cabinet Office policy setting out 18 criteria to 'help government create and run good digital services'.[68] Within these criteria, there are guidelines on collecting personal information from users, laying emphasis on the need to minimise the personal data that is collected to what is strictly needed for the provision of government services.[69] The GDSS principles are, to an important extent, aligned with the legal requirements in the GDPR and the Data Protection Act 2018. They are phrased, however, in very broad terms. They shall, thus, not be taken to impose an additional legal obligation to public authorities, particularly since there is no specific provision that would prohibit data collection akin to the one presently discussed. Nonetheless, it would be helpful if the Justice Data Infrastructure clarified its conformity with the GDSS within the public interest mandate for research data sharing we have proposed in this section.

65. European Data Protection Board, 'Opinion 3/2019 concerning the Questions and Answers on the interplay between the Clinical Trials Regulation (CTR) and the General Data Protection regulation (GDPR) (art. 70.1.b))' (23 January 2019) edpb.europa.eu/sites/edpb/files/files/file1/edpb_opinionctrq_a_final_en.pdf 8.
66. Lyon et al (n 14) 13.
67. *Infra* IV.
68. GOV.UK, 'Digital Service Standard' gov.uk/service-manual/service-standard
69. GOV.UK, 'Collecting personal information from users' gov.uk/service-manual/design/collecting-personal-information-from-users

02

Preparation and Linkage.

In this section:

1. Resources and Operating Model
2. Data Preparation
3. Data Linkage

Data Preparation & Linkage

This section discusses the key challenges in preparing data that has been collected for use by researchers. **First**, we discuss the preliminary issue of deciding the most appropriate overarching operating model for a data infrastructure, considering the impact of specific proposals on resource allocation and policy-making priorities within the HMCTS. **Second**, we present the main preparation principles that will allow for the effective use of administrative data for research purposes and reflect on their implementation in this context. **Third**, we outline the main alternative data linkage frameworks to maximise the analytical potential of the data infrastructure's resources.

KEY PIECES OF LEGISLATION FOR THE PREPARATION AND LINKAGE STAGE

- *EU General Data Protection Regulation (GDPR)*
  - » *Articles 5(1)(e) and 5(1)(f)*
- *European Convention on Human Rights (ECHR)*
  - » *Article 8*
- *Charter of Fundamental Rights of the European Union (EUCFR)*
  - » *Articles 7 and 8*

1. Resources and Operating Model

The creation of a bespoke data preparation and linkage framework for a new data infrastructure is a challenging process. It requires delicate balancing between achieving efficient data use and complying with the law. The benefits arising from the more effective and efficient use and re-use of data do not detract from the need to comply with legal requirements. Even if the potential to improve service delivery and reduce costs in the long-term is recognised, operational realities, including financial constraints and the challenges of policy-making, can create obstacles to investing the resources necessary to ensure the best outcome. For example, the Academy of Medical Sciences documented in 2011 the existence of a culture within the healthcare setting that 'fails to fully support the value and benefits of health research'.[70] Beyond the healthcare setting, the Data Sharing Review Report, published in 2008, stated that its most important recommendation was the improvement of the personal and organisational culture of those involved in data sharing. It acknowledged that information sharing carries both benefits and risks but that a 'culture of indecision' and risk aversion was 'problematic' and needed to change.[71]

A key distinction in this area is between a ***demand-led*** and a ***supply-led*** data preparation model.[72] The overarching operating rationale has significant implications for the resources required to establish and sustain any data infrastructure. In the case of a demand-led model, datasets are prepared for research use *in response* to a particular research request and tailored to the specific needs of the project. In contrast, in a supply-led model, core datasets are created, curated and maintained within government departments, allowing researchers to express an interest in accessing them.[73]

70. The Academy of Medical Sciences, '*A new pathway for the regulation and governance of health research*' (2011) acmedsci.ac.uk/file-download/35208-newpathw.pdf 7.
71. R Thomas and M Walport, '*Data Sharing Review Report*' July 2008 https://webarchive.nationalarchives.gov.uk/+/http:/www.justice.gov.uk/docs/data-sharing-review.pdf 54.
72. The two models are also framed as 'create-and-destroy' and 'retain-and-reuse' respectively, Research Project Management, 'Administrative Data Research Network (ADRN) Mid-Term Review Report' (8 November 2016) esrc.ukri.org/files/research/research-and-impact-evaluation/adrn-mid-term-review-report-8-november-2016 22. After the end of the research project, datasets are normally destroyed in the first case and retained in the second.
73. *Ibid*.

We advocate here for the latter approach, which we believe allows Government Departments to target resources towards the analysis of datasets that they believe might benefit them the most. The Administrative Data Research Network (ADRN) provides an interesting case study about the impact of different operating models on a data centre's function:

THE ADRN – FROM A DEMAND TO A SUPPLY-LED MODEL

In its first phase of operation (2013-2016), the ADRN attempted to collaborate with government departments on the premise of a demand-led model. Research requests would first be approved by the Network's Approvals Panel and then a request for the creation and linkage of bespoke datasets would be submitted to data providers such as the Department of Work and Pensions (DWP) or the HMRC.

In its fourth year of operation, the ADRN had managed to acquire data for approximately 19% of the approved projects it had been facilitating.[74] According to its Mid-Term review report, one of the main reasons for these data acquisition difficulties was the resource-intensive nature of the demand-led model. The result of such an approach was to foster counter-incentives for data providers.[75] In response, the ADRN implemented a major overhaul of its operating model and adopted a supply-led approach.

The idea behind the overhaul has been that a 'themed' approach to data acquisition, i.e. the identification of core research themes around which the datasets would be prepared, demonstrates more compellingly the value of data and the potential benefits for government departments.[76] Though this precludes the creation of bespoke datasets for individual research projects, the improvements in timelines, data access, and certainty were viewed as sufficiently beneficial as to warrant a demand-led approach.[77]

The ADRN's successor, the Administrative Data Research Partnership (ADRP), has strategically adopted a similar focus on 'prioritised policy themes'.[78] The eight core strategic themes are: housing and communities, health and well-being, children and young people, world of work, growing old, inequality and social inclusion, climate and sustainability and crime and justice.[79]

The demand- and supply-led models are not necessarily mutually exclusive and, thus, it would be possible for a data infrastructure to adopt a hybrid policy. This could entail the creation of a number of core datasets whilst also accepting requests for bespoke datasets on a case-by-case basis. Nonetheless, the potential of a primarily supply-led approach to clarify the expectations of policy-makers about the return of their investment in a data infrastructure should not be underestimated.

74. UKSA, 'Thirteenth Meeting of the ADRN Board: Agenda and Papers' (2017) https://perma.cc/K9TP-WL7A
75. *Supra* (n 71) 21.
76. K Jones et al, '*The Good, the Bad, the Clunky: Improving the Use of Administrative Data for Research*' (2019) 4(1) International Journal of Population Data Science 7.
77. ESRC, 'Administrative Data Research UK' https://perma.cc/VN8Z-PW4X
78. ESRC, 'Administrative Data Research Partnership' esrc.ukri.org/research/our-research/administrative-data-research-partnership
79. ADR UK, 'What are the ADR UK's main areas of research?' https://www.adruk.org/our-research

First, such a model stimulates demand around a set of research themes that serve the core aims of the HMCTS reform, e.g. efficiency and accessibility of the justice system. High-quality research conducted on the pertinent datasets would directly feed into the policy-making process, enabling a more a robust evaluation of a given policy choice.

Second, it would allow an identification of organisational actors which could collaborate with HMCTS in delivering the service and performing e.g. data de-identification and linkage. This is common in international best practice, with existing data infrastructures building valuable partnerships with trusted actors that can utilise their technical skills at the data preparation and linkage stages of the data processing activity. Such partnerships help to stabilise expectations as to the required investments and the potential returns of the data infrastructure. The range of potential synergies here would also rely on our proposed classification of data infrastructure users,[80] potentially allowing for the contribution of non-public bodies under certain conditions. Third, as it will be shown in the following sections, curation and retention of core datasets will allow the streamlining of preparation and linkage processes.

2. Data Preparation

Data preparation needs stem from the fact that administrative data is routinely collected without a subsequent research use in mind, making it a 'blunt measure for most theoretical constructs of interest'.[81] Even if some of this report's recommendations regarding data quality at the collection stage are adopted, this may not prevent issues arising at the data preparation and linkage stage. To achieve the data infrastructure's envisaged purposes, datasets must be prepared according to a set of agreed-upon information quality standards that will contribute to the validity and reliability of research using justice data.[82]

International best practice in this context suggests the FAIR principles are a key instrument for facilitating the preparation and linkage of disparate datasets. These principles have not only been endorsed by supranational and intergovernmental organisations such as the EU Commission and the G20,[83] but also by a wide number of stakeholders from the academic community and major research funders.[84]

80. *Supra* I.1
81. D DeHart and C Shapiro, *'Integrated Administrative Data & Criminal Justice Research*' (2016) American Journal of Criminal Justice.
82. C Reimsbach-Kounatze, '*The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis'*, (2015) OECD Digital Economy Papers, No. 245 22.
83. EU Commission, 'Turning Fair into Reality' (2018) publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283
G20, 'Leaders' Communiqué: Hangzhou Summit' (5 September 2016) https://perma.cc/N6T9-45DS
84. M Wilkinson et al, '*The FAIR Guiding Principles for scientific data management and stewardship*' (2016) Scientific Data 3;
Wellcome Open Research, 'Data Guidelines' wellcomeopenresearch.org/for-authors/data-guidelines

The principles can be defined as follows:

THE FAIR DATA PRINCIPLES [85]

Findable *– findability refers to the deposition of data in a stable and recognised repository, assigned a unique persistent identifier to allow discoverability by both humans and machines. The appropriate use of metadata can be instrumental in this respect.*[86]

Accessible *– accessibility refers to the use of such user licenses as the CC0 license and the OSI-approved license that facilitate data and source code re-use respectively.*[87] *Note that accessibility does not necessarily require open data sharing under all circumstances; ethical or confidentiality considerations can be observed within accessibility policies.*[88]

Interoperable *– interoperability refers to the adoption of common operational standards that enables systems to exchange and make use of data from different sources. The adoption of a standard vocabulary and file formats is highly desirable.*

Reusable *– reusability relies on the three former guiding principles and additionally encourages the inclusion of documentation alongside the data that will facilitate it being understandable and thus reusable.*

85. *Ibid* 'data guidelines'. .
86. M Boeckhout et al, '*The FAIR guiding principles for data stewardship: fair enough?*' (2018) 26 European Journal of Human Genetics 931.
87. Creative Commons, 'CC0' creativecommons.org/share-your-work/public-domain/cc0 Open Source Initiative, 'Licenses & Standards' opensource.org/licenses
88. Wellcome Open Research, 'Policies – Data Availability' wellcomeopenresearch.org/about/policies#dataavail

This is not to say that the FAIR principles should exclusively inform the data infrastructure's preparation strategy. Nonetheless, they provide a clear structure that presents significant overlaps with other examples of best practice:

Overlaps in Information Quality Principles:

FAIR	OECD	HIQA
Findable	**Timeliness, Relevance, Accuracy**	**Timely Dissemination**
Accessible	**Accessibility**	**–**
Interoperable	**Interoperability**	**Comparability**
Reusable	**Coherence, Credibility**	**Systematic Evaluation of Data Quality**

Despite these overlaps, some dimensions of data quality are not covered by the FAIR principles yet are crucial for present purposes. First, the OECD [89] guiding principles of relevance, accuracy, and coherence hint at the challenge of preparing datasets to ensure that the contents of these datasets (rather than their format or structure) are apt for research purposes. Second, the FAIR principles describe a 'process for accessing discovered data' and thus do not touch upon on moral, ethical or legal requirements that may be demanded in a particular data sharing context.[90]

While the FAIR principles aim to maximise the use of research data for knowledge discovery purposes, legal and ethical considerations may, at times, indicate that a maximal approach to data re-use is not appropriate. This is also dictated by such legal principles as data minimisation and data security, as well as by the requirement for proportionate interference with data subjects' rights to privacy and data protection.[91] Hence, the FAIR principles provide a helpful starting point, whilst also requiring further elaboration through the creation of context-dependent principles in response to the particular challenges of any given data environment.

Information governance standards such as the ones enunciated by the Health Information and Quality Authority (HIQA)[92] are helpful in this regard. These principles indicate that best practice requires the creation of a set of concrete data preparation policies that respond to the bespoke needs and novel characteristics of a Justice Data Infrastructure. A Jisc report on the application of the FAIR principles in practice examined the way in which data analysts and other officials handling administrative datasets perform actions required by the principles. This work identified not only a lack of understanding by the relevant officials around data ownership, management, and machine readability, but also an absence of supporting infrastructure for effective data management and metadata capture.[93] These organisational realities strongly undermine what a data infrastructure is trying to achieve with the FAIR principles within its data preparation policy. If the technical infrastructure or the expertise to capitalise the benefits of making data FAIR are missing, much of the importance of introducing the principles in the first place is diminished.

89. OECD, 'Quality Framework and Guidelines for OECD Statistical Activities' https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm
90. B Mons et al, '*Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud*' (2017) 37 Information Services and Use 49.
91. GDPR arts. 5(1)(e) and 5(1)(f); ECHR art 8; EUCFR arts 7 and 8.
92. Health Information and Quality Authority, '*Information management standards for national health and social care data collections*' https://www.hiqa.ie/sites/default/files/2017-02/Information-management-standards-for-national-health-and-social-care-data-collections.pdf
93. R Allen and D Hartland, '*FAIR in practice - Jisc report on the Findable Accessible Interoperable and Reuseable Data Principles*' (2018) zenodo.org/record/1245568#.XNwh4NNKgQ 6-7.

These challenges highlight the need to understand how the FAIR principles will operate in practice, a process that depends upon individuals within an institution having the requisite expertise, and the creation and maintenance of the most appropriate technical infrastructure. Such needs must be considered in light of the resources available to the data infrastructure to ensure the most effective and efficient management of these processes. In terms of expertise, the data infrastructure's analysts must be able accurately to assess the quality of information within a dataset, perform the appropriate cleaning of these datasets, and prepare them for matching.

With regard to information quality standards, considering that information that is 'good' for administrative use may be unsuitable for research purposes, it is likely that decisions will have to be made in the data collection stage and subsequently inform the data infrastructure's overarching policy. This will also depend on the extent to which relevant standardisation can cover the existing datasets or whether it will only be used for newly created datasets throughout the digitalisation process. Furthermore, HMCTS analysts will need to develop and update a metadata architecture, i.e. a set of tags and encoding schemes that will be used to describe the data infrastructure's resources.[94] This architecture will need to be designed with a view to ensure interoperability with the e-Government Metadata Standard[95] and other departmental metadata architectures used in key target datasets held by other data providers.

In respect of technical infrastructure, a preliminary question concerns clarifying whether there is convergence between the organisational structures involved in the administration of justice and the information and communications technology utilised by HMCTS. Deficiencies in infrastructure can hamper both the findability of accurate information and the interoperability of datasets that the data infrastructure aspires to integrate. Italian Trial Online (TOL), the Italian information system designed to provide access to procedural documents and notifications and allow payment of fees in civil cases, demonstrates the potential issues that can arise in this context.[96] In its initial version, there was no reliable way to verify the legal validity of documents, resulting in a failure to stimulate its widespread use.[97] Considering the sensitivity and gravity of the data that the data infrastructure will seek to link and share for research purposes, a trustworthy infrastructure that addresses any potential verification concerns is critical. Finally, differences in infrastructure between the data infrastructure and other data providers are likely to result in low interoperability, challenging the ability of the data infrastructure to link and integrate data effectively.[98] Whilst these considerations will require the investment of time and resources, clarity on the data infrastructure's overarching operating model, they will ultimately contribute to the creation of a sustainable data resource with significant potential to generate valuable insights in the long term. The final sub-section of this chapter elaborates on challenges in linking justice data.

94. HMRC, 'Departmental Metadata Architecture' webarchive.nationalarchives.gov.uk/20090903102800/http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=1033
95. Cabinet Office, 'e-Government Metadata Standard' www.nationalarchives.gov.uk/documents/information-management/egms-metadata-standard.pdf
96. G Lupo and J Bailey, '*Designing and Implementing e-Justice Systems: Some Lessons Learned from EU and Canadian Examples*' (2014) 3 Laws 353, 357.
97. *Ibid*.
98. Referring here to semantic interoperability, i.e. the ability of different information systems to 'communicate information consistent with the intended meaning of the encoded information', R Moser et al, '*Grid-Enabled Measures*' (2011) 40 American Journal of Preventive Medicine S134.

3. Data Linkage

Data already held by the HMCTS or generated through the digitalisation process could give unique and beneficial insights into societal problems when integrated with datasets held by other providers. Hence, a distinct set of issues within data preparation, concerns strategy with regard to ***data linkage***, i.e. the processes involved in 'connecting records that relate to the same person, family, event, organisation, or location within or between datasets'.[99] Linkage also happens *within* datasets for purposes of data cleaning, concerning the removal of duplicate records and verification of entity identities, but our present focus is on linkage between datasets for purposes of data integration and combined analysis.

As a process, linkage operates by reference to specific ***variables***, i.e. attributes that are recorded in both datasets of interest. It is successful when the variables of interest match in a record pair.[100] Prior to the linkage of different datasets, it is necessary to ensure that these datasets are compatible. This requires the data sources to have common and clear administrative coverage and time reference points to reduce the scope for coverage error[101] and multiple common variables to facilitate the validity of the record linkage.[102]

In matching records across different datasets, there is a need to balance the need for valid linkages with the risk of data subjects' identification on a systemic level. This concern informs the design of different linkage methodologies. A well-established method is ***person identifiable data–based (PID-based) linkage***, which is further distinguished into *deterministic* or *probabilistic* linkage.[103] ***Deterministic*** linkage is possible when there is a common unique identifier between two data sources, such as an individual's NHS number in NHS medical records. Agreement is thus determined on an 'all or nothing' basis, though the match status can be assessed in a single step or multiple step process. The single-step process compares the different records at once on the full set of identifiers whereas a multiple step process allows for approximate deterministic linkage by matching records through a set of progressively less restrictive steps. This approach classifies pairings between records as a match where they meet the criteria at any stage.[104]

Probabilistic linkage, on the other hand, involves the comparison of identifying variables across different datasets, aiming to estimate the probability of two records referring to the same individual. To do so, weights are assigned to specific variables with a margin of error from which an overall score is used to determine probabilistically matched pairs.[105] In theory, deterministic linkage is preferable due to the link between datasets being certain and simple to apply. The operational realities of data management, however, such as incomplete or erroneous input in records, often make probabilistic matching a more appropriate approach. Whilst there is a risk of inaccuracy, there has been sufficient methodological progress that probabilistic approaches can often achieve 'high, representative levels of complete linkage'.[106]

99. K Jones and D Ford, '*Privacy, Confidentiality and Practicalities in Data Linkage*' gss.civilservice.gov.uk/wp-content/uploads/2018/12/11-12-18_FINAL_Kerina_Jones_David_Ford_article.pdf 3.
100. *Ibid*.
101. Referring to a statistical bias that occurs when the target population does not coincide with the population actually sampled, see P Lavrakas, 'Coverage Error' (SAGE Encyclopedia in Survey Research Methods) https://methods.sagepub.com/Reference//encyclopedia-of-survey-research-methods/n115.xml for further information on this bias and its implications for statistical analysis
102. C Rao and M Kellya, '*Overview of the principles and international experiences in implementing record linkage mechanisms to assess completeness of death registration*' (May 2017) www.un.org/en/development/desa/population/publications/pdf/technical/TP2017-5.pdf S Dusetzina, S Tyree, A-M Meyer, et al An Overview of Record Linkage Methods in Linking Data for Health Services Researtch: A Framework and Instructional Guide (Agency for Healthcare Research and Quality, 2014) available at https://www.ncbi.nlm.nih.gov/books/NBK253312
103..Wellcome Trust, 'Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report' (March 2015) wellcome.ac.uk/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf
104. *Supra* (n 101).
105. C Rao and M Kellya, '*Overview of the principles and international experiences in implementing record linkage mechanisms to assess completeness of death registration*' (May 2017).
106. J Bentley et al, '*Investigating linkage rates among probabilistically linked birth and hospitalization record*' (2012) 12 BMC Medical Research Methodology 149.

The ultimate decision as to which form of linkage to use will depend upon the quality and nature of the data held by the respective data controllers.

Alternatively, linking methods that do not use person identifiable data for record comparison have been developed. These are framed as ***privacy-preserving record linkage*** (**PPRL**) methods since they involve hash-encoded records that are non-identifiable, refraining from extracting any identifiable data from the original dataset.[107] While this is a very appealing method from an information governance perspective by reducing privacy risks,[108] it is also quite demanding in terms of record accuracy and adequacy. As Wellcome Trust notes, the most obvious drawback is that it excludes the ability to use probabilistic linkage. Therefore, whilst two different records like 'JohnSmith' and 'JBSmith' could be probabilistically matched,[109] the application of a PPRL method would entail these records being represented by entirely different hash-encoded, non-informative identifiers.[110] With PPRL methods being less well-established than PID-based linkage, we assume that the data infrastructure will often be restricted to using identifiable data at the linkage level. This is consistent with the approach taken by existing data infrastructures. For example, the Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (RDC-IAB) finds that 'the loss of information due to the necessary anonymisation would be too great' and prefers to apply stricter disclosure controls when providing access to data.[111]

Furthermore, international best practice in this area indicates that existing data infrastructures have been using the services of *trusted intermediaries* with experience in securely de-identifying and linking datasets. The main example of this model involves a so-called *trusted third party* (TTP) system. A TTP is an 'independent organisation that acts as a liaison between two or more collaborating' organisations.[112] Such an intermediary organisation is responsible for the promotion of mutual trust between data sharing collaborators and the prevention of re-identification of data subjects throughout the process. The TTP employs encoding procedures and matching algorithms, first de-identifying the relevant records from different sources by assigning anonymous identifiers to them and then allowing their integration within a databank or data infrastructure. There are numerous examples of data infrastructures using TTPs to manage information risks and comply with legal requirements. In the Netherlands, the Ministry of Security and Justice used ZorgTTP (a Dutch TTP specialising on health care applications) which employed a double, one-way hashing procedure to integrate addiction treatment and resettlement data to estimate the number of problem drug users.[113] Similar examples of successfully employing TTPs are the Centre for Data Linkage in Australia,[114] Population Data BC in Canada,[115] ONS in England[116] and NWIS in Wales.[117]

107. *Supra* (n 101) 65.
108. Especially in cases where the data are not allowed to be transferred out of the DataLab and have to be anonymized at source, *ibid*.
109. *Ibid*.
110. A Brown et al, '*Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets*' (2017) 17(1) BMC Med Inform Decis Mak 83.
111. M Antoni and A Schmucker, '*The Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (RDC-IAB)*' (2019) 4(2):5 IJDPS.
112. S W Van den Braak et al, '*Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector*' (2012) Proceedings of the 13th Annual International Conference on Digital Government Research
113. *Ibid* 142.
114. *Supra* (n 37) 13.
115. P Hertzman et al, '*Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest*' (2013) 20(1) J Am Med Inform Assoc 25-8.
116. ONS, 'ONS policy for safeguarding data whilst managing Admin Data Research Network projects' https://perma.cc/MB5M-3W3Q
117. K Jones et al, '*SAIL Databank: 10 years of Spearheading Data Privacy and Research Utility*' (2017) saildatabank.com/wp-content/uploads/SAIL_10_year_anniversary_brochure.pdf

A clear articulation of the main stages of a TTP operation can be identified by reference to the work that NWIS performs for the SAIL databank in Wales:

MAIN STAGES OF DATA DE-IDENTIFICATION AND LINKAGE IN THE SAIL DATABANK [118]

1. Splitting the datasets *– datasets are first split into a* ***demographic*** *(e.g. name, date of birth, gender) and clinical (e.g. medication records and procedures) component.*[119] *The demographic component is transferred to the TTP, whereas the clinical component goes to the Databank using a web-based secure file transfer system.*

2. Anonymisation and encryption *– the TTP anonymises and encrypts demographic records, assigning an Anonymous Linking Field (ALF)*[120] *to each one of them.*

3. Re-combining the datas*ets – anonymised and encrypted demographic elements are sent to the Databank, containing only the ALF, week of birth, gender code, and area of residence. These elements are then integrated with the clinical components and are now able to be linked with other datasets.*

4. Additional safeguards *– the Databank performs further encryption of the ALF to form an ALF-E, which is used to perform linkage across datasets. Further safeguards are applied when linkage with a small dataset is desired to avoid any possibility of identification.*

Identifying an appropriate TTP for the Justice Data Infrastructure will be informed not only by existing relationships between HMCTS and relevant organisations, but also by the data mapping exercise we have described in the previous section. Some TTPs may have more experience in linking specific types of data or have a pre-existing relationship with specific data providers, whose datasets might be a linking target for the Justice Data Infrastructure.

While the use of a TTP does not pre-empt information risks in subsequent data processing cycles,[121] its adoption at the linkage stage would lay the foundations for minimising such risks to an acceptable level, in line with legal requirements and best practice.[122] The following section discusses the infrastructure's approach to *providing access* to collected, prepared, and potentially linked datasets to researchers.

118. SAIL Databank, 'The Anonymisation Process' saildatabank.com/saildata/data-privacy-security/#anonymisation-process

119. The clinical component is an example of content component; in the HMCTS case, we could have a separation of demographic data from such content component as the outcome of a judgment or a settlement.

120. ALFs are unique identifiers assigned to each individual represented in a dataset that replace the commonly recognised identifiers of a particular dataset. For example, SAIL creates ALFs based on a person's NHS number which is then encrypted using a Blowfish algorithm.

121. On a context-specific assessment of identification risks by reference to 'data environment', see M Mourby et al (n 37) 222.

122. GSS, 'Privacy and data confidentiality methods: a National Statistician's Quality Review' (13 December 2018) gss.civilservice.gov.uk/policy-store/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review-nsqr

03

Data Access by Researchers.

In this section:

1. A Governance Structure for Data Access
2. Strategic Facets of a Research Data Access Policy
3. Operational Aspects of Data Access by Researchers

Data Access by Researchers

Data collection and preparation are prerequisites to the data infrastructure's central purpose: providing researchers with access to data, enabling the evaluation and analysis of the justice system. This section tackles some of the key challenges around data access. **First**, we propose the implementation of a governance structure to address strategic and operational challenges arising from data access. **Second**, we delineate the core strategic facets of a research data access policy, particularly the definition of public interest and the main principles of engagement with the public and data users. **Third**, we present the main considerations around ensuring proportionate data access to the data infrastructure's resources, drawing on operational legal compliance and ethical compliance requirements.

KEY LEGISLATION AT THE ACCESS STAGE

- *EU General Data Protection Regulation (GDPR)*
  - » *Articles 4(5) and 89*
  - » *Recitals 28 and 156*
- *Digital Economy Act (DEA) 2017 UK*
  - » *Sections 64-70*
- *Data Protection Act (DPA) 2018 UK*
  - » *Section 8*
- *Statistics and Registration Services Acts (SRSA) 2007 UK*
  - » *Section 7*
- *European Convention on Human Rights (ECHR)*
  - » *Article 8*
- *Charter of Fundamental Rights of the European Union (EUCFR)*
  - » *Articles 7 and 8*

1. A Governance Structure for Data Access

Prior to the establishment of a data infrastructure, certain questions concerning operations and governance must be tackled. One example in the previous section is the decision regarding whether the HMCTS should adopt a 'demand-led' or 'supply-led' operating model. This section concerns another fundamental operational aspect; the *governance structure* that will determine the who, what, where, why, and how researchers may access data held by HMCTS.

A governance structure operates not only to ensure that data is accessed by researchers in compliance with all relevant legal regulations, but also that access is in line with the values and standards upon which HMCTS is founded. Currently, the responsible body to perform these functions is the Data Access Panel (DAP).[123] At the moment, DAP is an 'email group supported by a small secretariat and leadership function in the Analysis and Performance team'.[124] The group does not record minutes of its meetings, as there is no legal or business requirement to do so.[125] Similarly, information on the data requested by researchers or the outcome of requests is not published and, even when legally requested via a FOIA request, is often heavily redacted.[126] There are reasons for believing that the current capacity and operation of the DAP would not satisfy governance requirements in the advent of increased research interest in justice data and HMCTS' commitment to streamline access processes to facilitate such research.

123 . HMCTS, 'Access to courts and tribunals for academic researchers' www.gov.uk/guidance/access-to-courts-and-tribunals-for-academic-researchers

124 . Byrom (n 4) 8.

125. Ministry of Justice, 'Freedom of Information Act (FOIA) Request – 181127019' (27 December 2018) www.whatdotheyknow.com/request/535603/response/1285629/attach/3/FOI%20181127019%20Phil%20Booth%2027%20December%202018%20Final.pdf?cookie_passthrough=1 3.

126. *Ibid*.

Starting with the legal conditions imposed in this context, recent legislative developments have somewhat clarified the legal basis upon which certain data sharing activities can take place between public bodies and third parties. The Digital Economy Act (DEA) 2017 introduces a generic legal power for public bodies to share data for research purposes subject to compliance with the UK Data Protection Act (DPA) 2018.[127] This 'gateway' to data sharing is conditioned upon compliance with the requirements set out in section 64 DEA. These include, *inter alia*, the de-identification of data to ensure that individual identities are 'not reasonably likely to be deduced', steps to avoid 'accidental or deliberate' data disclosure, accreditation of both the requesting researcher and the research project by the UK Statistics Authority (UKSA) Board and compliance with the Code of Practice established by the UKSA Board under s. 70 DEA.[128] Considering that designing data sharing agreements and performing data disclosure to researchers may be beyond the scope of many officials' knowledge base, a clear and well-known governance structure is integral to managing access requirements. Good governance, generally, requires an 'accessible articulation of the different values and standards against which individual and organisational activity will be assessed'.[129] Such values and standards inspired by the HMCTS reform mission,[130] as well as the requirements in the DEA and the UKSA Code of Practice need to work in harmony. An oversight body is necessary to ensure this is feasible, ensuring that HMCTS officials are well-informed about their powers and duties, and confident in their decisions to allow access to the data infrastructure's resources for research use.[131]

There are also practical, day-to-day access-related challenges that the governance structure will have to address. Two main tasks are the oversight of data sharing agreements between HMCTS and researchers and the enforcement of such agreements through the imposition of the appropriate disciplinary measures. Similar infrastructures, including the HMRC datalab, require researchers to sign a service level agreement and condition continuous access to data upon their compliance with its terms.[132] This practice is in line with the legal requirements regarding appropriate safeguards to mitigate data processing risks in EU data protection law. Article 89 of the GDPR imposes this requirement when data is used for research purposes, with the ICO recommending the use of particular techniques such as encryption and anonymisation.[133]

Constant oversight of researchers to ensure they adhere to the terms of the data sharing agreements and the adopted safeguards is, however, very resource intensive. A more effective approach would be for the governance structure to agree upon the use of a range of technical and organisational measures that are operative before any agreements are made and regularly assess the adequacy of these measures. Other initiatives have confined data sharing to their own secure physical spaces[134] or require specific researcher training and accreditation to reduce the risk of accidental or negligent harmful disclosures.[135] In case researchers are non-compliant with the prescribed requirements, the governance structure should be able to impose disciplinary penalties. Potential penalties are outlined in the UK Statistics Authority's Code of Practice, which draws its legislative authority from s70

127. s. 65(2)(a) DEA.

128. For a more comprehensive discussion of the requirements, see J Bell et al, '*Balancing Data Subjects' Rights and Public Interest Research: Examining the Interplay between UK law, EU human rights law and the GDPR*' (2019) 1 EDPL 43, 46.

129. N Sethi and G Laurie, 'Delivering proportionate governance in the era of eHealth: Making linkage and privacy work together' (2013) 13 Medical Law International 168.

130. Ministry of Justice, 'Transforming Our Justice System' assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/553261/joint-vision-statement.pdf

131. S Timmis et al, 'Sharing the benefits: How to use data effectively in the public sector' reform.uk/research/sharing-benefits-how-use-data-effectively-public-sector 3.

132. HMRC, 'Research at HMRC' www.gov.uk/government/organisations/hm-revenue-customs/about/research

133. ICO, 'Guide to the General Data Protection Regulation' (September 2018) ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf 24.

134. E.g. the HMRC DataLab, see M Almunia et al, '*Expanding access to administrative data: the case of tax authorities in Finland and the UK*' (2019) 26 Int Tax Public Finance 661-676.

135. E.g. the ONS using the UKSA's Researcher Accreditation Panel, see ONS, 'Accessing secure research data as an accredited researcher' www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme#becoming-an-approved-researcher-through-the-ons-approved-researcher-scheme

DEA, and include the loss of researcher accreditation and the termination of access to data. The emphasis on disciplinary penalties, rather than formal criminal sanctions, arises from the OECD's findings that such penalties are more effective than formal criminal sanctions in this context.[136]

Whilst the establishment of a governance structure to manage data access undoubtedly aligns with best practice in this area, there is still a need to carefully consider the most appropriate characteristics and functions of such a structure. This is contingent not only on the requirements imposed by law and the principles and values central to the objectives of HMCTS but also on the available financial and human resources within the system. Nevertheless, there are two generic considerations that can assist policy design. First, to avoid the duplication of effort and resources, we recommend that HMCTS identify relevant actors that operate within the space of research data sharing and consider involving them in its initiatives. Such an exercise would offer helpful sources of guidance and examples of best practice that have been proven to work 'on the ground.' A prime example is the UKSA-based National Statistician's Data Ethics Advisory Committee (NSDEC).[137] This committee was originally established to advise the National Statistician on the ethical appropriateness of research projects and policy proposals that wished to access data from the Office for National Statistics (ONS) and the Government Statistical Service (GSS). NSDEC are integral to ensuring transparency around the access, use, and sharing of data for research and statistics by elucidating the aspects of a proposal that may raise ethical issues in its nuanced, multi-faceted consideration of the appropriateness of a proposal.[138]

In doing so, it considers more than the relevant legal requirements and as its decisions are also guided by considerations as to the social acceptability of the proposal. NSDEC's expertise and experience in this area is widely acknowledged, as exemplified in the approach of the Administrative Data Research Network's Approval Panel. Whilst a separate body, the ADRN Approvals Panel refers applications to the NSDEC so that the latter can 'provide ethical consideration for government and third sector researchers wishing to access data via the ADRN,'[139] in light of its extensive capabilities. We recommend that the HMCTS draws upon the experience of NSDEC in determining the appropriateness of data access proposals in an advisory capacity.

Second, an effective governance structure operates to engender the values that support and promote an organisation's identity and to ensure that its processes are respected as authoritative. Such an approach both facilitates the greater sharing of data but also, through aligning with the existing values of an organisation, promotes the 'buy-in' of key internal stakeholders. A combination of the capacity to represent diverse stakeholders and the competence to make decisions based on professional experience and expertise is instrumental to this purpose. Whilst it is beyond the scope of this report to definitively determine whether this is best achieved through a management team, a steering committee, or an independent group of external advisors,[140] valuable insights may be drawn from existing best practice on the duties and powers of similar governance structures.

136. OECD, 'OECD Expert Group For International Collaboration On Microdata Access: Final Report' (2014) www.oecd.org/std/microdata-access-final-report-OECD-2014.pdf
T Desai et al, Five Safes: designing data access for research www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf
137. UKSA, National Statistician's Data Ethics Advisory Committee www.statisticsauthority.gov.uk/about-the-authority/committees/nsdec
138. UKSA, Data Ethics www.statisticsauthority.gov.uk/about-the-authority/committees/nsdec/data-ethics
139. UKSA, NSDEC Terms of Reference https://perma.cc/LXF9-J33N
140. E Welch et al, '*Institutional and Organizational Factors for Enabling Data Access, Exchange and Use in Genomics Organizations*' (2016).

DATA ACCESS GOVERNANCE STRUCTURES IN SAIL AND THE ADR UK

The Information Governance Review Panel (IGRP) is designed to manage the complex issue of data access by researchers within SAIL. The IGRP comprises of information governance and ethics experts drawn from a variety of stakeholder bodies (e.g. the British Medical Association, NHS Wales Informatics Service) and members of the public. It is tasked with ensuring that 'that the limited data put together for researchers conforms to Information Governance regulations'[141] *and overseeing researcher compliance with the terms of SAIL data sharing agreements. The SAIL team benefit from the IGRP's feedback on whether approving a disclosure request serves the 'public interest' and how perceived risks in linking the requested data can be mitigated.*

The Research Commissioning Board (RCB) and the Operational Management Group (OMG) are groups with similar responsibilities within the ADRP. The RCB and the OMG have taken up the core access-related responsibilities of the now defunct ADRN Approvals Panel. The RCB assesses the conformity of research proposals with the 'public interest', aiming to identify new opportunities for administrative data research. Its members are external to ADR UK and are drawn from diverse backgrounds, … and a range of organisations'.[142] *The OMG is responsible for 'monitoring and reporting of all operational matters of ADR UK',*[143] *including assessing researcher compliance with the terms of ADR data sharing agreements.*

Drawing on the ways in which other data infrastructures have designed their data access governance structures is an excellent resource for guiding the design of a Justice Data Infrastructure. The following section will elaborate on the core strategic facets of a research data access policy that the infrastructure's governance structure will be called to shape.

2. Strategic Facets of a Research Data Access Policy

A number of overarching issues should guide the data infrastructure's general approach to data access, beyond the specificities of a certain proposal's compliance with the various legal and ethical requirements. We here provide a discussion of three core issues: defining the 'public interest', engaging with the general public, and the Justice Data Infrastructure's users.

141. R Lyons et al, '*The Secure Anonymised Information Linkage (SAIL) system in Wales has privacy protection at its heart*' (2014) BMJ.
142. ADR UK, 'Governance' www.adruk.org/about-us/governance
143. *Ibid*.

A. Defining the 'Public Interest'

This discussion builds on our recommendation for a principled classification of users by reference to the public interest. The public interest is often portrayed as an 'elusive' concept [144] that risks being defined by reference to whatever an institution wants to do, rather than what it should do.

In the British public-sector, what comes closer to an authoritative definition of the 'public interest' for data-intensive research purposes is to be found in the statistical research context.

THE 'PUBLIC INTEREST' IN STATISTICAL RESEARCH LEGISLATION

'Public interest', often used interchangeably as 'public benefit' or 'public good', is defined in section 7 of the Statistics and Registration Service Act (SRSA) 2007 as: 'informing the public about social and economic matters; assisting in the development and evaluation of public policy; and regulating quality and publicly challenging the misuse of statistics'. This definition is to be read in conjunction with the provisions of the UKSA's Research Code of Practice and Accreditation,[145] which was initially a voluntary code of best practice developed by the English Office for National Statistics and became legally binding under section 70 of the DEA 2017. The Code's criteria reflect the inclusive nature of the 'public interest' definition, that can promote a wide range of important research and facilitates the ability of researchers to demonstrate their work satisfies the Code's requirements. More specifically, the Code requiress that the 'primary purpose' of the proposed project is:

an evidence base for 'public policy decision-making (…), public service delivery (…) or decisions which are likely to significantly benefit the economy, society or quality of life of people in the UK' or

to 'replicate, validate, challenge or review existing research and proposed research publications' or

to 'significantly extend understanding of social or economic trends or events by improving knowledge or challenging widely accepted analyses' or

to 'improve the quality, coverage or presentation of existing research, including official or National Statistics'.

144. S King et al, '*Reflections on Defining the Public Interest*' (2010) 41(8) Administration & Society;
D MacNair, '*Government Lawyers and the Elusive Concept of Public Interest: A Canadian Perspective*' in P Keyzer (ed), *Public Sentinels*.
145. UKSA, 'Research Code of Practice and Accreditation Criteria' (1 March 2018) www.gov.uk/government/consultations/digital-economy-act-part-5-data-sharing-codes-and-regulations/research-code-of-practice-and-accreditation-criteria

Both NSDEC, in developing their own concept of 'public benefit' to assess a project's compliance with any relevant ethical requirements,[146] and the ONS, within their '5 safes' framework of assessing legal, moral and technical appropriateness of data sharing,[147] have drawn upon this articulation of the public benefit.

Beyond the statistical research context, a Justice Data Infrastructure could benefit from drawing upon the practices of other initiatives that have elaborated the notion of the public interest. The Micro-Data Release Panel (MRP), part of the HMRC's datalab, have adopted a narrower definition of the public interest that aligns with the research interests and policy agendas of HMRC and HMT.[148] Valuable guidance may also be drawn from initiatives that have considered the notion of the 'public benefit' in the context of the wider sharing of health data for research purposes. Understanding Patient Data have proposed a set of overarching principles to ensure conformity of a data sharing request with the public benefit.[149]

First, data sharing should be ***purposeful***: the purpose should be 'clear and transparently defined', explaining the tangible benefits for both individual patients and the potential for improving public health as a social good. **Second**, disclosure should be ***proportionate***: the minimum amount of personal data to achieve the proposed goal should be disclosed under clear conditions and upon a holistic consideration of the risks arising from disclosure. **Third**, data sharing should be ***responsible***: secure and effective use of the data should be guaranteed in the interest of delivering the intended outcomes.

How does, however, a data infrastructure practically ensure that the perceptions of its own managing / steering committee about the public good correspond to those of the wider public?

B. Public Engagement

It is common for lay representatives to be included within the governance structures of similar initiatives. This serves not only to demonstrate compliance with the principle of transparency and as tangible evidence of respect for the public interest, but also as a core strategic pillar of an institution's overall access policy. Policy Connect, a cross-party think-tank, published a report that highlighted the importance of public engagement in the use of public sector data.[150] In this report, they caution against the creation of rules or policies with 'little or no public engagement (…) which could contribute to public distrust in data use'.[151] They propose a variety of alternatives for meaningful engagement with the public, including 'open consultations, town-hall meetings, industry outreach, and other ways of directly engaging with members of the public and relevant stakeholders'.[152] Policy Connect see a by-design, strategic commitment to public engagement as instrumental to maintaining public confidence and acceptability of public sector data sharing.

146. NSDEC, 'Guidelines on using the ethics self-assessment process' https://perma.cc/LXF9-J33N
147. Desai (n 136).
148. Welpton and Wright (n 12) 5.
149. Understanding Patient Data, 'Data for public benefit : balancing the risks and benefits of data sharing' www.involve.org.uk/sites/default/files/field/attachemnt/Data%20for%20Public%20Benefit%20Report_0.pdf
150. Policy Connect, 'Trust, Transparency and Tech' www.policyconnect.org.uk/appgda/sites/site_appgda/files/report/454/fieldreportdownload/trusttransparentcyandtechreport.pdf
151. *Ibid* 15.
152. *Ibid*

This assertion is supported by empirical findings in the existing literature. A deliberative public engagement event in British Columbia, Canada yielded the conclusion that, provided that adequate data privacy and security safeguards are in place, patients are supportive of streamlining data access procedures for research 'because of the value it provides to society'.[153] This suggests that public scepticism about big government data research may have more to do with lack of knowledge about the strict data security safeguards that are applied and less with an inherent hostility against collecting and retaining personal information on a big scale.

Public engagement is integral to consolidating the trustworthiness of HMCTS in respect of robust protection of data privacy and facilitate public support for the data infrastructure's research data sharing aspirations. The link between meaningful engagement with the public and the societal acceptability of research data sharing is also supported by a study of public engagement experience in the context of the Scottish Health Informatics Programme.[154] Focus group participants indicated a preference for transparency and openness on behalf of data researchers, rather than a reiteration of the many positives that can come out of data-intensive research.[155] The findings of another empirical study into trust between stakeholders of administrative data research in England are similar: more transparency about the data sharing purposes and processes employed by a data infrastructure resulted in increased public trust in the appropriateness of data sharing.[156] A helpful example of a comprehensive public engagement policy comes from ADR UK:

153. J Teng et al, '*Sharing linked data sets for research: results from a deliberative public engagement event in British Columbia, Canada*' (2019) 4(1) IJDPS.
154. M Aitken et al, '*Moving from trust to trustworthiness: Experiences of public engagement in the Scottish Health Informatics Programme*' (2016) Science & Public Policy.
155. *Ibid*.
156. A Sexton et al, '*A balance of trust in the use of government administrative data*' (2017) 17(4) Archival Science 305, 324.

PUBLIC ENGAGEMENT WITHIN ADR UK

The ADRN investment, and its successor ADR UK, perceive the role of public engagement as central within their governance frameworks. ADR Wales and ADR Scotland both maintain public panels with an advisory capacity on the alignment of their projects with the public interest.[157]

In Wales, the ADR Consumer Panel for Data Linkage Research was established in 2011 and inputs on the Centre's governance frameworks, public engagement policies, and research practices. ADR Wales management have found the panel's contributions to their work very valuable, with the views of its members providing a 'positive outlook and a fresh, and sometimes unexpected, perspective on various issues'.[158] *In Scotland, the ADR Public Panel for Scotland similarly guarantees the participation of the wider public in the shaping of the Centre's frameworks and policies.*

Beyond these initiatives, which include the views of the public through individual lay representation, the ADR engage with Voluntary, Community and Social Enterprise (VCSE) organisations to understand the research needs of 'marginalised groups and empower these groups by offering the opportunity to influence the direction and outcome of research'.[159] *In 2017, ADR Northern Ireland hosted a Data Workshop series in partnership with Detail Data – a VCSE organisation – to connect NGOs with administrative data research. This initiative aimed to make such data more 'community-relevant' and raise awareness of the potential power of public data among different communities.*

Drawing on these examples of best practice, we recommend the Justice Data Infrastructure should strive for a meaningful protection of the public interest through engaging with the wider public. Such engagement would ensure public trust in the research undertaken on HMCTS data. A meaningful engagement with researchers, i.e. its users, is another step of strategic importance in this direction.

C. User Engagement

Beyond ensuring that the voices of the wider public are duly considered within its data access policy, the data infrastructure should also consider the needs of researchers accessing its resources. Implementing in practice a coherent definition of users and ensuring that they are well-trained to use the infrastructure's resources is integral to complying with legal and ethical requirements of data access for research purposes.

157. ADR UK, 'How do we work with the public?' www.adruk.org/our-mission/working-with-the-public
158. *Ibid*.
159. *Ibid*; this would be particularly important in the context of a Justice Data Infrastructure, considering the likely vulnerability of various users accessing the justice system.

Having already discussed the issue of defining 'users',[160] we now turn to the practical implications of adopting such a definition for a data access policy.

Best practice suggests the data infrastructure, through its governance structure, would have to engage in ongoing evaluation of the types of entities or individuals that are eligible to apply for data access. This is the case regardless of the particular user definition adopted. Some data infrastructures, e.g. the Justice Data Lab at the English Ministry of Justice, allow access by VCSE organisations, social enterprises and private businesses,[161] whereas ADR UK allows access only by academic researchers.[162] In other cases, different types of users may be allowed access, albeit only under different conditions. The Virtual Micro-data Laboratory (VML) at the English ONS, for instance, allows direct access for government users, but requires academic researchers to obtain an 'Approved Researcher' accreditation by contacting the UK Data Service and satisfying the ONS's relevant requirements.[163]

Following the ONS's initiative to establish such an accreditation framework, it would be fruitful for the HMCTS to reflect on the relative merits in either creating a framework for 'HMCTS accredited' researchers, or strive for the establishment of a common accreditation framework across government. Regardless of the particular contents of such an accreditation framework, it is crucial that its requirements are applied in a consistent, transparent, and accessible manner, in the interest of creating and maintaining reasonable expectations about the data infrastructure's access policy.

Reasonable expectations around access matter not only when it comes to the wider public to bolster support for research use of justice data, but also as far as various relevant stakeholders in the justice system, e.g. justice-oriented social researchers, VCSE researchers or government analysts, are concerned. To the extent that the data infrastructure's requirements are perceived as fair and justifiable, it is more likely that all interested parties will feel inclined to pursue justice data research and enhance the long-term success prospects of HMCTS' investment. Furthermore, the consistent and transparent practical implementation of a principled definition is instrumental to an iterative process of evaluating and, potentially, revising such a definition. By experiencing the practicalities of engaging with different users, the data infrastructure's team will acquire rich insight into the different benefits and risks that each entity created when they accessed justice data. Are different users complementing each other or are they duplicating work?

In addition to the practical implementation of a user definition, the Justice Data Infrastructure would need to carefully consider the appropriate training of its users. Looking at best practice in this area, it is apparent that researcher training schemes ensure that users possess high levels of relevant knowledge, the skills to use the available resources and are incentivised to use the data in an appropriate manner. Due to the legal requirement under s. 70 of the DEA 2017 to comply with the training requirements stipulated in the UKSA's Research Code of Practice, it would be helpful to have regard to the organisational practice from which these requirements originated, i.e. the ONS's 'Approved Researcher' scheme:

160. *Supra*, I, 1-5 where three distinct ways of classifying the infrastructure's users were discussed and a classification around the notion of 'public interest' was endorsed.
161. Lyon et al (n 14).
162. ADR UK, 'How do we work with researchers?' www.adruk.org/about-us/our-partnership/adr-northern-ireland
163. ONS (n 22).

THE ONS 'APPROVED RESEARCHER' SCHEME

Proceeding from a generic legal requirement in the SRSA 2007 to develop criteria for appropriate data access for statistical research purposes, the English ONS, through its Microdata Release Panel, developed a sophisticated researcher accreditation system.[164]

Individuals who want to access the VML's resources need to possess the relevant knowledge and skills that the ONS has elucidated through a set of key requirements. ***First****, the aspiring 'approved researcher' must have a qualification (at least an undergraduate degree) including a significant proportion of maths or statistics. In the alternative, they must demonstrate 'at least 3 years quantitative research experience'.*[165] ***Second****, the applicant must undergo the Safe User of Research data Environments (SURE) training course, administered by the ONS, the UK Data Service, the ADR UK or HMRC. In this training course, researchers are encouraged to consider how core aspects of data security, legal and ethical compliance apply to their proposed research. They also learn through experience by evaluating numerous examples of (fake) outputs as 'safe' or 'unsafe' for disclosure. The course ends with a comprehensive examination.*[166] ***Third****, researchers must agree to a set of publicity requirements, i.e. their inclusion to an accredited researchers list published on the ONS website, the publication of results, and adherence to a formal accredited researcher declaration.*

The ONS also provide for 'provisional accreditation' in case an individual lacks the qualifications or the statistical experience required, subject to meeting a set of other criteria.[167]

164. Note that this system has two prongs: 'approved researcher' and 'approved research'.
165. ONS (n 22).
166. V Moody, '*A new integrated approach: training researchers to use sensitive microdata*' (June 2016) blog.ukdataservice.ac.uk/a-new-integrated-approach-training-researchers-to-use-sensitive-microdata
167. ONS (n 22).

The SURE training course has been adopted by other initiatives, including ADR UK.[168] The particular needs of a Justice Data Infrastructure will determine whether it can be utilised in this context or whether a bespoke training scheme might be preferable. HMRC opted for the latter option in the context of the HMRC datalab. This course is delivered by HMRC officials and consists of four modules: 'introduction to the DataLab', 'keeping data safe', 'statistical disclosure control' and 'bookings and outputs'.[169] The aims of this course are to familiarise researchers with the legislation applicable to HMRC and explain how the datalab operates, including procedures on how to request outputs and the rules of the IT room. After two years, researchers need to undertake a refresher course. Potentially, HMCTS could combine elements of an existing and well-recognised training scheme like SURE with bespoke features that are important in the justice context e.g. courses on the relevant legislation or the particularities of justice data. Nonetheless, beyond these strategic issues, the Justice Data Infrastructure will need to cope with day-to-day, operational challenges around proportionate disclosure. The following section elaborates on these challenges.

3. Operational Aspects of Data Access by Researchers

By *operational* aspects, we refer to day-to-day access to the Justice Data Infrastructure's resources. *Proportionate* disclosure of data lies at the heart of maximising the potential of data to improve our knowledge about the justice system, whilst guaranteeing the appropriate protection of the rights of data subjects. We first discuss the legal dimension of proportionality and then the need to observe ethical requirements throughout the process of data access.

168. UK Data Service, 'Access to the Secure Lab' www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab/train
169. Welpton and Wright (n 12) 6.

A. Legal compliance

The rights to privacy and data protection are, in human rights law terms, *qualified* rights, i.e. they may be interfered with when this is necessary, yet only when such an interference is proportionate to the legitimate aim pursued. On a conceptual level, proportionality implies some type of balancing between competing considerations.[170] In our case, the rights of data subjects, whose personal data is held by the data infrastructure, need to be balanced with the public interest in the use (and re-use) of justice data for evaluation and analysis purposes. Chapter I has already presented some of the key legal issues in respect of the data infrastructure's collection strategy, i.e. 'lawful and fair processing' and 'purpose limitation and data minimisation.[171] The difficulties with relying on consent to use data in research, were discussed there.[172]

In principle, to achieve proportionate access, one proceeds from a privacy-by-design approach to inform the various layers of the infrastructure's governance. Privacy-by-design may have various implications when considering the relevant data protection and human rights law requirements stipulated in domestic legislation (DPA 2018, DEA 2017), EU legislation (the GDPR), and human rights law (ECHR, EUCFR).[173] For instance, it might necessitate anonymising all available data by default, requiring researchers to make the case for the necessity of identifiable data for their projects.[174]

170. K Möller, '*Proportionality and Rights Inflation*' in G Huscroft et al (eds), Proportionality and the Rule of Law: Rights, Justification, Reasoning (CUP 2014).
171. *Supra*, I, 8-12
172. Sethi and Laurie (n 129).
173. See Bell et al (n 128) for a comprehensive analysis of 'appropriate safeguards' under article 89 GDPR and their interplay with human rights law.
174. Sethi and Laurie (n 129).

On the other hand, and in the interest of avoiding excessive barriers to data sharing, it is important to ensure that the regulatory burden does not greatly outweigh the relative risks through using privacy and data protection impact assessments. Balancing privacy/data protection with the public interest in improved knowledge about the justice system requires a case-by-case, detailed assessment of what is at stake, acknowledging that 'different degrees of protection, sharing, oversight and, ultimately sanction' will be needed from time to time.[175] Nonetheless, it is only at the time of actual access by a researcher or a group of researchers that all these measures are really put to test.

How, then, can a data infrastructure apply the most appropriate controls to govern researcher access to its resources?

Existing best practice suggests that when complete anonymisation is not possible, ***pseudonymisation***, also known as ***de-identification***, of data is integral to achieving a proportionate solution. While the precise meaning of pseudonymisation hinges on how one defines 'identifiability' and is contested in the literature, [176] the GDPR definition will be the focus of present discussion:

'PSEUDONYMISATION' AND PROPORTIONATE RESEARCH DATA ACCESS

In the ICO's words, pseudonymisation is 'the process of distinguishing individuals in a dataset by using a unique identifier which does not reveal their 'real world' identity.[177]

Article 4(5) GDPR provides a more comprehensive definition of pseudonymisation as 'the processing of personal data in such a manner that they can no longer be attributed to a specific data subject without the use of additional information'.

This definition assumes that additional information 'is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'.

Pseudonymisation is particularly relevant to proportionate data access in the context of data sharing for research purposes, considering that it is explicitly mentioned in the GDPR (Article 89 and Recitals 28 and 156) as an 'appropriate safeguard' to reduce identification risks for data subjects in the context of scientific, historical or statistical research.

175. Sethi and Laurie (n 129).
176. Mourby et al (n 37); cf. M Berberich and M Steiner '*Blockchain Technology and the GDPR – How to Reconcile Privacy and Distributed Ledgers?*' (2016) EDPL 424
177. Information Commissioner's Office, '*Anonymisation: managing data protection risk code of practice*' (Wilmslow, November 2012) ico.org.uk/media/1061/anonymisation-code.pdf

Different models have been adopted across data infrastructures and similar projects in the interest of pseudonymising their datasets and, thus, work towards upholding proportionality in data access by researchers.

The SAIL databank, whose sophisticated data linkage security safeguards were discussed in the previous chapter, have created a robust management system to handle dataset information and researcher access: the *SAIL Gateway*.[178] After establishing feasibility of a request, the SAIL team refer it to their Information Governance Review Panel (IGRP) for an information governance review before granting access. The Panel highlights perceived identification risks, informing, in turn, the SAIL team to guide preparations for researcher access. Crucially, through the Gateway, access is *remote* via a secure system protected by firewalls, two-factor authentication processes, password-protected servers, and encrypted network connections. Researchers log in via a Virtual Private Network (VPN) and an authentication token which generates a one-time password when placed in the computer's USB slot. Researchers are prevented from misusing data by copying or transferring information that they are not allowed to since the SAIL system controls via software the configuration of the remote desktop. The SAIL team have also received an external verification of their information governance compliance by inviting an independent internal audit.

Other data infrastructures have developed similar control processes. The ONS Virtual Microdata Laboratory provides on-site access to 'approved researchers', prohibiting unauthorised removal of information from the premises.[179] It seeks to strike the right balance between data security and researcher access by placing strict security arrangements on the one hand,[180] while allowing researchers to combine different sources of data within the VML's secure confines on the other. The UK Data Service Secure Lab prevents unauthorised access by either confining access to on-site within a Safe Room if the data is very sensitive or, in any case, using a secure encrypted web-based interface that prevents data download.[181] The HMRC datalab pseudonymises administrative tax records, allowing access only on-site, within a safe room.[182] Population Data BC, a multi-university data access and linkage health research resource in Canada,[183] first pseudonymises the research data extracts and then uses a secure environment that can be accessed from anywhere in Canada to grant access to authorised researchers. Legal compliance itself, however, is not sufficient. The data infrastructure will also have to address ethical questions when reviewing access requests.

B. Ethical requirements

A data access policy underpinned by general principles of proportionality and engagement with the general public also needs to identify and mitigate the ethical risks of disclosure on a case-by-case basis. In most cases, data infrastructures seek the recommendation of external bodies that have relevant experience and specialise in deliberating on the reasonable expectations of the public from the use of their data in research. NHS Digital, for example, treats information governance compliance and satisfaction of ethical requirements as distinct layers of governance. The Confidentiality Advisory Group (CAG) advises the Health Research Authority (HRA) when access to patient data without consent for research use is being sought.

178. D Ford et al, '*The SAIL Databank: building a national architecture for e-health research and evaluation*' (2009) BMC Health Services Research https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-9-157
179. ONS (n 22).
180. ONS (n 22); heavy penalties (such as custodial sentences in the examples of the VML and Secure Data Service) can be imposed upon researchers who abuse their access rights.
181. UK Data Service (n 168).
182. Almunia et al (n 134) 674.
183. Hertzman et al (n 115) 27.

From an ethical review perspective, researchers need to apply to a Research Ethics Committee (REC) and seek its approval of the proposed project.[184] The use of RECs in the UK reflects a broader international practice which emerged in response to unethical research practices, aiming to protect the best interests of research participants and the broader public, as well as cultivate trust and confidence in public-interest research.[185] There are more than eighty (80) NHS RECs across the UK and their membership consists of up to fifteen (15) members, a third of whom are 'lay'.[186] To assess whether a proposed research project is ethical, Committees process applications through a robust and thorough procedure,[187] involving on-going oversight throughout the research project and granting applicants the right to appeal and challenge their verdict. To maintain their status as impartial, RECs are 'entirely independent of research sponsors, funders and the researchers themselves'.[188]

Closer to the subject matter of a Justice Data Infrastructure , the example of the National Statistician's Data Ethics Advisory Committee (NSDEC) is relevant and, potentially, of direct assistance, as we have already suggested when discussing the allocation of governance tasks to other organisations.[189] The NSDEC advises the National Statistician on the ethical appropriateness of using ONS and GSS data for research by reference to identifying 'clear benefits for users and (…) the public good'.[190] Proportionality is the cornerstone of this assessment, since benefits to individuals or society do not suffice if the proposed research is not *necessary* to realise them.

Countervailing considerations include the potential harms that could arise from data disclosure and their proportionality to these benefits. Potential harms do not only relate to reidentification risks, but also to the risk of perpetuating biases via data sources, algorithms, and presentation of research outcomes.[191] The proposed methods, population coverage and aspired aims are rigorously assessed in that regard. While data subjects' consent is not legally required in all cases, the Committee places particular emphasis on the circumstances under which consent is relied upon from an ethical perspective. Power imbalances between research participants and researchers, or public authorities which may be seen as endorsing research, render such an assessment crucial. The NSDEC also consider the risks raised by the introduction of new technologies in research, assessing whether methods employed are consistent with recognised standards of integrity and quality. Finally, yet importantly, the Committee aim to raise the awareness of researchers in respect of their project's potential ethical risks by providing them with a self-assessment form 'for them to review the ethic of their projects'.[192] This form highlights the risks of linking sensitive data, including the data of children and vulnerable adults, urging researchers to take appropriate action and amend their proposals accordingly before seeking the Committee's approval. A Justice Data Infrastructure would benefit significantly from the methods employed by the NSDEC and the RECs, complementing them with the context-specific considerations that apply to justice data. We conclude this report by discussing the infrastructure's retention and re-use policy.

184. HRA, 'Research Ethics Service and Research Ethics Committees' www.hra.nhs.uk/about-us/committees-and-services/res-and-recs
185. C Thomson, '*Research Ethics Committees*' (2012) Encyclopedia of Applied Ethics 786.
186. HRA, 'Research Ethics Committees Overview' www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/research-ethics-committees-overview
187. HRA, 'Research Ethics Committee – Standard Operating Procedures' www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/research-ethics-committee-standard-operating-procedures
188. HRA (n 186).
189. *Supra* 36.
190. UKSA (n 139).
191. UKSA (n 137).
192. UKSA (n 137).

04

Data Retention and Re-Use.

In this section:

1. Opting for a Retain-and-Reuse Model
2. Designing a Retention Policy
3. Key Elements of the Justice Data Infrastructure's Research Retention-and-Reuse Policy

Data Retention and Re-Use

The final chapter addresses *data retention* and *re-use*, i.e. the Justice Data Infrastructure's overall approach to sustaining data resources and stimulating research interest in them over time. **First**, we clarify that adopting a 'retain-and-reuse' model is consistent with the aspirations of the MoJ and HMCTS to facilitate independent research on the reform programme. **Second**, we outline the legal opportunities and constraints that are present in designing a data retention policy for the Justice Data Infrastructure. **Third**, we draw on international best practice to suggest what the core facets of such a policy should be, focusing on the potential risks of long-term identifiability of data subjects and transparent re-use of the Justice Data Infrastructure's resources.

KEY PIECES OF LEGISLATION FOR THE RETENTION AND RE-USE STAGE

- *EU General Data Protection Regulation (GDPR)*
  - » *Articles 5(1)(b), (c), (e) and (f), 85 and 89*
- *Data Protection Act (DPA) 2018 UK*
  - » *Sections 170, 171, 174(3), 174(3)(a), 176(1), sch. 2 part 6 para. 27 and sch. 2 part 5 para. 26(2)(b)*
- *European Convention on Human Rights (ECHR)*
  - » *Article 8*
- *Charter of Fundamental Rights of the European Union (EUCFR)*
  - » *Articles 7 and 8*

1. Opting for a Retain-and-Reuse Model

Unlike the three previous stages of data processing, i.e. collection, preparation and access, a comprehensive data retention and re-use approach is not a necessary element of creating a research data infrastructure. As discussed in the second chapter of this report,[193] such research initiatives as the ADRN preferred a 'create-and-destroy' model. Under this model, individual requests for the creation and linkage of bespoke datasets were submitted to data providers and data would be deleted after the completion of particular research projects. We have already made the case for a predominantly 'supply-led' model, which would entail the creation of a number of key datasets and the stimulation of research interest in them.[194] We argued that this would help to serve the core aims of HMCTS reform, i.e. a more efficient and accessible justice system, better than a 'demand-led' model being driven by individual researchers' requests.

A data retention-and-reuse approach is, similarly, justified by reference to realising these aims. In one of the reform updates made public in May 2018, the CEO of HMCTS committed to making data, 'in a suitably anonymised way', available for researchers and academics to use.[195] This commitment was elaborated upon in later HMCTS publications, where it was clarified that 'independent research' on the reform programme is integral to establishing accountability and transparency.[196] To achieve this, HMCTS will need to improve the way they share data with external researchers, who, as recent empirical work demonstrates, are facing a number of difficulties in accessing justice data.[197]

193. *Supra* chapter 2.
194. *Ibid.*
195. HMCTS Chief Executive, 'Modernising the Courts and Tribunals Service: Future of Justice Conference' (14 May 2018) https://perma.cc/N6N2-3AC3
196. Ministry of Justice, 'Evaluating our reforms: Response to PAC Recommendation 4, January 2019' www.gov.uk/government/news/moj-response-to-public-accounts-committee-transforming-courts-and-tribunals
197. Byrom (n 4) 31.

With not enough information about the data held and made available by HMCTS, such researchers are dependent on developing relationships with supportive individuals in HMCTS to gain access.[198]

The design and publication of a comprehensive retain-and-reuse policy is integral to alleviating uncertainty about the data infrastructure's resources. It would also indicate the key types of data that researchers could rely on gaining access to without excessive delays.[199] This would promote long-term partnerships with external researchers and making long-term access sustainable, in the interest of achieving dynamic, on-going and independent evaluation of the impact of the reform programme on the users of the justice system.[200]

However, the idea of retaining and constantly expanding massive datasets, which may include individual personal data, raises both legal and ethical challenges.[201] In the advent of increased public and private data-driven surveillance over the last few decades, it is understandable that the combination of the massive amount of data and the prospect of its indefinite storage and re-use could create public concern.[202] Such a policy, therefore, cannot be designed without reference to key governance and legal considerations that will ensure legitimacy and public acceptability.

2. Designing a Retention Policy

Is it lawful to retain and update large datasets, which partly store identifiable data, with a view to re-using them for research purposes over a long time? If yes, what are the legal requirements for doing so? The answer to these questions is lies in the delicate interplay between data protection and human rights law. We discuss these two areas of law to identify the requirements that a retention and re-use policy in a Justice Data Infrastructure will need to adhere to.

A. Data Protection Law Principles

From the perspective of data protection law, the question of retention is to be placed within the broader interplay between the relevant data protection principles and the exemptions made in the legal framework for academic research (re-)use. Article 85 GDPR allows Member States to derogate from the Regulation in the interest of reconciling the protection of personal data with the right to freedom of expression, 'including processing for (…) the purposes of academic (…) expression'.[203] While the UK has extensively exercised this capacity to establish a comprehensive regime of exemptions in the Data Protection Act 2018,[204] the scope of such exemptions may not always be as wide and encompassing as often envisaged.[205]

198. *Ibid*.
199. This is not to suggest, then, that all delays in facilitating researchers' access to HMCTS data are unjustified since appropriate information governance standards require that careful consideration of the controls applied to minimise risks for data protection is undertaken.
200. HMCTS Chief Executive (n 195).
201. The assumption being here that a Justice Data Infrastructure will be storing personal data in the interest of maximising the analytical potential of future research; see the similar approach of a Data Centre at the German Federal Employment Agency, Antoni and Schmucker (n 111).
202. Particularly if the public think that commercial enterprises might be allowed access to the data at some point, see IPSOS, 'The One-Way Mirror: Public Attitudes to Commercial Access to Health Data' (March 2016) www.ipsos.com/sites/default/files/publication/5200-03/sri-wellcome-trust-commercial-access-to-health-data.pdf 18.
203. Article 85 GDPR.
204. Data Protection Act 2018 (UK) (hereafter cited as 'DPA'), ss 170 and 171.
205. M Mourby et al, '*Governance Of Academic Research Data Under The GDPR – Lessons From The UK*' (2019) 9(3) IDPL 192.

This raises the need for considering the interplay between principle and exemption in the particular context:

KEY PRINCIPLES FOR DATA RETENTION

While all six data protection principles apply, the four most challenging requirements from a data retention perspective are ***purpose limitation***, ***data minimisation***, ***storage limitation*** *and* ***data security***.[206]

Purpose limitation*, under article 5(1)(b) GDPR, allows the processing of data for a purpose other than the one for which it was collected only where the new purpose is 'compatible' with the original one.* ***Data minimisation****, under article 5(1)(c) GDPR, dictates that data processing shall be limited by reference to requirements of relevance and necessity 'in relation to the purposes' of such processing.*

Storage limitation*, under article 5(1)(e) GDPR, prescribes that data are stored in an identifiable form only 'as long as necessary'.* ***Data security****, or 'integrity and confidentiality' under 5(1)(f) GDPR, requires the existence of 'appropriate technical or organisational measures' that will safeguard personal data from both unauthorised interference and accidental damage.*

In the first instance, there appears to be a tension between the requirement for a minimalist and strictly-purposeful processing of data and the inherently ambitious endeavour of establishing a comprehensive Justice Data Infrastructure for research re-use.[207] Furthermore, the more ambitious such an infrastructure is, e.g. by allowing on-going remote access to researchers, the more demanding it becomes from a data security perspective.[208] This need for 'appropriate safeguards' from a data security perspective is also articulated in article 89 GDPR when data is processed for statistical research purposes.

206. A Tamò-Larrieux, *Designing for Privacy and its Legal Framework* (Springer 2018) 11.
207. See D Erdos, 'Systematically Handicapped? Social Research in the Data Protection Framework' (2011) 20 ICTL 83 on the 'fluid and norm-challenging nature of a social science research endeavour' and its tension with data protection principles.
208. Cf. the HMRC Datalab's approach which only allows on-site access, Almunia et al (n 134) 674.

The research exemptions established by the GDPR and the DPA 2018 mitigate the rigour of these requirements to ensure that they do not 'prevent or seriously impair' the achievement of the purposes of processing,[209] i.e. in our case the facilitation of independent research on the HMCTS reform programme. Both purpose and storage limitation provide for an explicit exception if the data is held and re-purposed for 'scientific research' purposes,[210] and international best practice has developed effective approaches to satisfying data minimisation and security safeguards.[211] Nonetheless, these exemptions are far from unconditional. For these exemptions to apply, research purposes must be the *sole* purpose of data processing, the ICO having the power to determine whether this is indeed the case.[212] The processing should always be linked with the publication of academic material[213] and HMCTS need to have a reasonable belief that such a publication would be in the public interest.[214] The latter requirement indicates the previously discussed[215] significance of the creation of a clear and transparent 'public interest' mandate by HMCTS in its Justice Data Infrastructure-related policies and guidance. Provided that scientific research is the sole purpose of processing, it is not necessary to set a specific retention period and data can be kept for longer.[216]

209. DPA schedule 2, part 6, paragraph 27
210. Also see *supra* chapter I.
211. *Infra* section 3.
212. DPA s 174(3)(a).
213. *Ibid*, ss. 174(3) & 176(1).
214. *Ibid*, Schedule 2, Part 5, para 26 (2)(b).
215. *Supra* chapter I.
216. ICO, 'Principle (e): Storage limitation' ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/storage-limitation

B. Human Rights Principles

Even if an ambitious retention and re-use policy seems compatible with data protection law, under the discussed conditions, there is still a need to consider its conformity with human rights norms, i.e. the fundamental rights to privacy and data protection in the ECHR and the EU charter. The principle here is that long-term or indefinite retention of personal data needs to satisfy the requirements of *proportionality* within both human rights law frameworks: a legitimate aim, relevance, necessity and a fair balance between the aim and the interference caused by the retention.[217] Relevant case law mostly relates to retaining biogenetic or communications data for purposes of crime prevention.[218] Both the CJEU (in *Digital Rights Ireland*) and the ECtHR (in *S and Marper*) have struck down legislation that enabled retention of a 'blanket and indiscriminate nature', affording excessive discretion to the authorities as to what is retained, when and for how long.[219] In both cases, the Courts held that fair balance between the 'very considerable' public interest of preventing crime and the fundamental rights of data subjects had not been struck.

217. *Gaughran's Application for Judicial Review, Re* [2015] UKSC 29 (28-33).
218. A Vedaschi and V Lubello, '*Data Retention and its Implications for the Fundamental Right to Privacy*' (2015) 20 Tilburg Law Review 14.
219. Joined Cases C293/12 and C594/12 *Digital Rights Ireland Ltd v Minister for Communications* [2014] 3 WLR 1607; S and *Marper v United Kingdom* (2009) 48 EHRR 50.

An even more instructive, for present purposes, example comes from a recent UK Supreme Court (UKSC) case:

THE UK SUPREME COURT JUDGMENT ON ADMINISTRATIVE RETENTION OF PERSONAL INFORMATION

*In R. (**on the application of C) v Secretary of State for Work and Pensions**,[220] the UKSC was asked to assess the lawfulness of a Department for Work and Pensions (DWP) policy concerning gender reassignment of its customers. Under this policy, the DWP's system would retain gender change data for 50 years after the individual's death. An individual brought claims under article 8 ECHR, complaining that their right to private life as a transgender individual is interfered with by this policy.*

The Court was convinced that the DWP had a legitimate aim in retaining the relevant records 'for the purpose of calculating entitlement to state retirement pension and the need to identify and detect fraud'.[221] In assessing whether a fair balance had been struck with the intrusion into the complainant's private life, the Court laid significant emphasis on the 'special procedures for restricting access to the records of customers who required extra protection' established by the DWP.[222] While a vast number of customers was catered for by the specific policy, only 'rarely' would front-line officers be allowed to access the relevant database. Hence, the interference was proportionate and not unlawful.

Within the confines set by the law, it will be for the Justice Data Infrastructure's retention policy to demonstrate that a proportionate approach is adopted, balancing the legitimate aim to facilitate independent research on the HMCTS reform programme with data subjects' rights. We now turn to some of the core elements that such a retention policy should be based upon.

3. Key Retention and Re-Use Policy Elements

International best practice suggests that data infrastructures often adopt long-term retention policies for their datasets, curating and updating them on an on-going basis to facilitate high-quality health and social-science research.

One example is the South Australia and Northern Territory (SA and NT) Datalink, which provides data linkage services to enable 'academics and policy makers to undertake research, policy, planning and evaluation'.[223] The datasets held by the SA and NT Datalink are updated annually or quarterly to ensure that they satisfy researcher needs.

220. R. (on the application of C) v Secretary of State for Work and Pensions [2017] UKSC 72.
221. *Ibid* (17).
222. *Ibid* (41).
223. M Schneider et al, 'Population Data Centre Profile: SA NT DataLink (South Australia and Northern Territory)' (2019) 4(2):8 IJDPS.

Similar policies of not only regularly updating, but also expanding their research data assets are adopted by the Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (RDC-IAB),[224] the Centre for Data and Knowledge Integration for Health (CIDACS) in Brazil,[225] the Canadian Primary Care Sentinel Surveillance Network[226] and the Centre for Health Record Linkage in New South Wales.[227]

These data centres have publicised their commitment to strictly observing legal and ethical requirements through following appropriate administrative and governance processes in accordance with privacy-by-design.[228] Considering its proximity in respect of applicable legal and ethical frameworks to a Justice Data Infrastructure, the example of the German Federal Employment Agency DataLab is instructive:

THE RESEARCH DATA CENTRE OF THE GERMAN FEDERAL EMPLOYMENT AGENCY AT THE INSTITUTE FOR EMPLOYMENT RESEARCH

The Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (RDC-IAB) has, since 2004, linked data gathered from employers, other administrative processes in the labour market, and survey data for research purposes. As the data include unique identifiers, it is classed by German data protection law as highly sensitive and its use by researchers is strictly regulated.

The RDC-IAB engages extensively with the research community to take its needs into account through monitoring demand for its datasets, conducting user surveys, and its data user workshops. Such interactions have led to the development and expansion of the datasets, tools, and services offered by the RDC-IAB and the number of data products has quadrupled over the past 15 years. It is not just the number of products that shave increased as the RDC-IAB has also established centres in the UK and US where researchers can access the data it holds.[229]

Research utilising the data provided by the RDC-IAB has led to changes in labour market policy, for example where research demonstrated the rigidity of the existing approach did not work for certain groups of unemployed individuals and thus advocated for a more flexible approach to be adopted.[230]

224. M Antoni and M Schmucker (n 111).
225 . B de Araujo Almeida et al, '*The Center for Data and Knowledge Integration for Health (CIDACS) An Experience of Linking Health and Social Data in Brazil*' (2019) 4(2):4 IJDPS.
226 . S Garies et al, '*Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction, and processing in Alberta*' (2019) 4(2):2 IJDPS.
227. K Irvine et al, '*Centre for Health Record Linkage: expanding access to linked population data for NSW and the ACT, Australia*' (2019) 4(2):7 IJDPS
228. A Boulle et al, '*Data Centre Profile: The Provincial Health Data Centre of the Western Cape Province, South Africa*' 4(2):6 IJDPS.
229. D Mueller and J Moeller, '*Giving the International Scientific Community Access to German Labor Market Data: a Success Story*' in N Crato and P Paruolo (eds), Data-Driven Policy Impact Evaluation (Springer 2018) 101.
230. *Ibid* 101.

Legal and ethical approval frameworks differ across jurisdictions.[231] There are, however, some focal areas where international practice on research data retention converges: the need to appropriately anonymise datasets to minimise long-term identifiability risks and the need for transparent decision-making about retention procedures. We discuss both, suggesting that a Justice Data Infrastructure can draw valuable lessons from international practice.

A. The Separation Principle

With legitimate retention requiring a proportionate trade-off between the public interest and data subjects' rights, it is very important to mitigate the identifiability risks arising for users of the justice system by the on-going research use of a Justice Data Infrastructure. The design of data infrastructures in that regard is informed by the *separation* principle, i.e. separating datasets and storing different segments of data in separate databases.[232] This is consistent with data minimisation requirements, keeping data that will be re-used in a less identifiable form, rather than erasing it.

The process of data separation is quite similar to the one of linking datasets through a trusted-third-party (TTP), as described in the second chapter of the present report.[233] Identifiable information, e.g. names, addresses and dates of birth, are separated from non-personal, content data such as a medical diagnosis or, in the case of a Justice Data Infrastructure, the outcome of a case by stage (settled, withdrawn, judgment issued) or the value of settlements and judgments.[234]

The two types of datasets are retained separately and strict data security safeguards are applied to the information that links them together for the purposes of a particular research project.[235] Such safeguards often involve the use of intermediary infrastructures like TTPs:

APPLYING THE SEPARATION PRINCIPLE IN THE SOUTH AUSTRALIA AND NORTHERN TERRITORY DATALINK

The SA and NT DataLink's Data Integration Unit (DIU) assists health data custodians with separating anonymised clinical and demographic datasets,[236] *using the following process:*

1. Data custodians provide the demographic datasets to the DataLink.

2. The DataLink creates Project Specific Linkage Keys (PSLK) and returns them to the data custodians.

3. The custodians attach PSLKs to anonymised content datasets and provide them to the researcher.

4. The researcher integrates analyses anonymised content data from many custodians using the linkage technology.

231. Although domestic laws of such other states as Brazil have been largely inspired by the EU GDPR, de Araujo Almeida et al (n 225) 4.
232. Tamò-Larrieux (n 206) 12.
233. *Supra* chapter 2.
234. Byrom (n 4) 25.
235. Irvine et al (n 227) 2.
236. Schneider et al (n 223) 4.

Applying the separation principle is integral to the proportionate retention and reuse of justice data for independent research. On the one hand, linking the separate datasets allows the potential benefits from sharing justice data to remain intact.[237] On the other, as was the case with the DWP's retention of gender reassignment data,[238] the fact that data analysts and researchers would not be able to access identifiable data mitigates re-identification risks. This is also consistent with the ICO's guidance on indefinite data retention for the exclusive purpose of scientific research, where 'pseudonymisation' is mentioned as an appropriate safeguard.[239]

B. Transparent Decision Making

Separating datasets to mitigate identification risks does not exhaust the stewardship obligations of data infrastructures with regard to research retention-and-reuse. International best practice suggests that when data is retained for research purposes, transparency mechanisms should exist to provide to the data subjects an understanding of retention-and-reuse-related decision-making. Data centres publicise their particular governance arrangements and retention policies, indicating which are the responsible groups or structures for deciding what is to be retained, for how long and for what type of research.

As pointed out when discussing the design of the data infrastructure's overarching governance structure,[240] this is less a matter of establishing a particular type of deciding body, and more a matter of better understanding and respecting reasonable expectations of data reuse. In some cases, responsibility for designing and updating a retention policy may rest with a high-level 'steering committee',[241] whereas other data centres may entrust this to more informal groups such as a 'small group' of research institute members with expertise in data-intensive health research.[242]

Responsible bodies for research data retention and re-use manage legal transparency requirements, e.g. providing privacy notices that inform the public about the ongoing use of their data for research,[243] as well as broader mechanisms of public engagement. Such data centres as the SAIL Databank and the Western Australia Data Linkage Branch maintain consumer panels comprised of members of the general public.[244] This facilitates on-going communication between data subjects and data custodians, as well as researchers, informing the former about the re-use of their data within various research projects.

237. E Morrow, '*Administrative Data: Misuse vs. Missed Use*' (2 January 2020) www.adruk.org/news-publications/news-blogs/administrative-data-misuse-vs-missed-use-133.
238. *Supra* (n 220).
239. ICO (n 216).
240. *Supra* chapter 3.
241. Schneider et al (n 221) 2.
242. de Araujo Almeida et al (n 223) 3.
243. K Jones et al, '*A Profile of the SAIL Databank on the UK Secure Research Platform*' (2019) 4(2):3 IJDPS 4.
244. *Ibid* 5.

05

Conclusion.

Conclusion

In concluding this report, we distil our key findings into a series of recommendations. These recommendations aspire to lay the groundwork for the governance of a Justice Data Infrastructure which will enable ongoing research and evaluation of the ambitious HMCTS reform programme. We also suggest a number of fruitful lines of enquiry to be pursued in future research.[245] The present report, aiming to offer an overarching governance blueprint, does not delve deeply into some peculiarities of justice data that could merit further interrogation. For example, it would be important to assess the extent to which justice data are different from other types of administrative data in respect of the constitutional background and the type of organisations that are involved in their collection and management e.g. the Judiciary and the HMCTS or external organisations. How does this particular constitutional and organisational context make a Justice Data Infrastructure different from other research data infrastructures? Other important lines of enquiry include the design of appropriate and effective public engagement methods that can adequately capture the different justice contexts about the use of justice data, as well as the particular implications of using justice data in AI development (e.g. reflecting on the need to balance the rich data that machine learning models use with the needs of data minimisation in the justice context).

Our recommendations are organised to reflect the four stages of data processing that structure this report. Combining our analysis of applicable legal frameworks and international best practice in designing public-sector data infrastructures for research and evaluation., these recommendations highlight governance considerations that will allow a Justice Data Infrastructure to stimulate interest and mobilise expertise to improve our knowledge about the justice system, while at the same time safeguarding the fundamental rights of data subjects, including the most vulnerable.

USER CLASSIFICATION

1. Access to justice data for research and evaluation purposes should be allowed by reference to the ***contribution*** *of the requesting party to the* ***production of knowledge that will improve policy-making in the public interest.***

2. HMCTS should ***publicise and consistently apply a transparent policy****, including specific criteria, for a project to be considered as serving the public interest, as well as on the* ***different requirements*** *that should be applied to* ***various users*** *(e.g. academic researchers, charities, start-ups, or established private sector companies).*

3. HMCTS should consider the relative merits in establishing a specialised ***accreditation framework for HMCTS 'accredited researchers'*** *based on the previously mentioned data access policy, as opposed to utilising an existing accreditation framework or striving for the establishment of a common research accreditation framework across the UK public sector.*

245. We thank our workshop participants for kindly steering our thinking towards some of the suggestions that follow here.

DATA COLLECTION

*4. HMCTS should produce an outward-looking **data catalogue** that identifies available datasets for research and evaluation use, with a view to encouraging interested parties to explore promising avenues of research.*

*5. Available datasets should be identified not only by reference to the **internal function of the tribunals** (e.g. routine case management data, claim outcome data, demographic and equalities data); the catalogue should also highlight the potential of **linking such datasets** with those belonging to other departments in the interest of addressing **broader systemic questions** (e.g. enforcement of judicial decisions or prevention of disputes from arising).*

*6. HMCTS should continue to work with **stakeholder networks** in respect of data that will need to be collected throughout and after the reform process, and develop an appropriate strategy to ensure that recommendations are implemented.*

DATA PREPARATION AND LINKAGE

*7. HMCTS should prepare, curate and maintain datasets for research and evaluation use with a **long-term vision**. HMCTS should strive for high data quality to maximise datasets' analytical potential, in accordance with the established international principles of findability, accessibility, interoperability and re-usability (**the FAIR principles**).*

*8. Other government departments that hold datasets of potential analytical interest to the HMCTS should continue to be involved in this process on an on-going basis, and **key organisational alliances** with them should be strengthened within an overarching data linkage strategy.*

*9. HMCTS should consider collaboration with a **trusted intermediary (or trusted-third-party) with experience in securely de-identifying and linking datasets**. Such an intermediary should be an independent organisation capable of promoting mutual trust between HMCTS and interested parties, as well as public confidence about the prevention of re-identification of justice system users.*

DATA ACCESS

*10. A **standing governance structure** should be responsible for handling **operational aspects of data sharing**, such as information governance, researcher training, and legal compliance, as well as feeding back to the HMCTS about **strategic priorities**. The Data Access Panel (DAP) could fulfil this role, provided that its present capacity and mode of operation are strengthened to satisfy increasing research interest in justice data.*

*11. HMCTS should consider the delegation of particular governance tasks, such as ethical approval of a request to use the data infrastructure's resources, to **existing stakeholders with considerably experience and expertise** (e.g. the National Statistician's Data Ethics Advisory Committee).*

*12. **Public engagement** should be a core facet of data access policy, in accordance with international best practice that suggests **transparency about the data sharing purposes and processes** employed by a data infrastructure as an indicator of appropriateness in data sharing.*

*13. **User engagement** should inform the development of data access policy, allowing **on-going evaluation of the prescribed strategic priorities** and the accepted types of **data infrastructure users**.*

DATA RETENTION AND RE-USE

*14. HMCTS should opt for a **retain and re-use policy**, adhering to the key legal requirement of **proportionate** use of justice data on a sustainable basis to minimise the risk of interference with data protection and privacy rights of the data subjects.*

*15. In accordance with international best practice, HMCTS should **pseudonymise** datasets, separating content data (e.g. the outcome of a case) from demographic data (i.e. identifiable data such as names, addresses and dates of birth).*

About the Project and the Authors

Unlocking the Potential of AI for English Law

This research forms part of the ***Unlocking the Potential of AI for English Law*** project led by Professor John Armour, University of Oxford. The project is funded by the Industrial Strategy Challenge Fund's (ISCF) Next Generation Services Research Programme and UK Research and Innovation (UKRI), and involves collaborations between researchers in the Oxford departments and faculties of Law, Economics, Computer Science, Education and the Said Business School.

Stergios Aidinlis is a final-year DPhil in Socio-Legal Studies candidate at the University of Oxford, Faculty of Law. His research, funded by the ESRC and the Onassis Foundation, theorises data sharing regulation in the British public sector, particularly regarding the disclosure of Government-owned, administrative data for research purposes.

Hannah Smith is a DPhil student based at the Centre for Health, Law, and Emerging Technologies at the University of Oxford, Faculty of Law. Funded by the ESRC, her thesis explores the secondary use of administrative data in research and the divergences between what is legally permissible and societally acceptable in this context.

Abi Adams-Prassl is a Senior Research Fellow and Associate Professor in the Department of Economics, University of Oxford. Her work focuses on developing new empirical tools to better understand public and private regulatory choices, from female empowerment to access to justice.

Jeremias Adams-Prassl is Professor of Law and Fellow of Magdalen College, Oxford. He is particularly interested in the impact of digitalisation on access to justice and the future of work.