A Blueprint for a Social Data Foundation

Accelerating Trustworthy and Collaborative Data Sharing for Health and Social Care Transformation

WSI White Paper #4

November 2020

Michael Boniface, Laura Carmichael, Wendy Hall, Brian Pickering, Sophie Stalla-Bourdillon & Steve Taylor

University of Southampton

Web Science Institute

About the WSI

The Web Science Institute (WSI) co-ordinates the University of Southampton's (UoS) world-leading, interdisciplinary expertise in Web Science, to tackle the most pressing global challenges facing the World Wide Web and wider society today. Research lies at its heart, positioning it as a leader in Web Science knowledge and innovation and fuelling its extensive education, training, enterprise and impact activities. The WSI is also UoS's main point of contact with The Alan Turing Institute, the UK's national institute for Data Science and AI, of which UoS is a partner university.

https://www.southampton.ac.uk/wsi/index.page

https://www.southampton.ac.uk/wsi/enterprise-and-impact/policy.page

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the University of Southampton, the Web Science Institute or its Board of Directors.

Web Science Institute
Building 32, Highfield Campus, University of Southampton, SO17 1BJ
wsi@soton.ac.uk

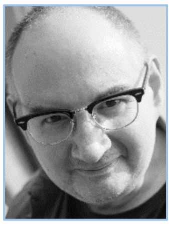
About the Authors

Michael Boniface is a Professorial Fellow of Information Systems in the School of Electronics and Computer Science, University of Southampton and Director of the University of Southampton, IT Innovation Centre. Michael has 20+ years' experience of applied research and innovation in federated systems management and data-driven innovation across healthcare, smart cities, engineering, telecommunications and creative industries. His digital health and health data research tackles challenges of data governance, data management and predictive analytics using AI/ML for population level health policies, emergency care decision support, and patient self-management of diseases.

Dame Wendy Hall, DBE, FRS, FREng is Regius Professor of Computer Science, Associate Vice President (International Engagement) and is an Executive Director of the Web Science Institute at the University of Southampton. She became a Dame Commander of the British Empire in the 2009 UK New Year's Honours list, and is a Fellow of the Royal Society and the Royal Academy of Engineering. Dame Wendy was co-Chair of the UK government's AI Review, which was published in October 2017, and is the first Skills Champion for AI in the UK. In May 2020, she was appointed as Chair of the Ada Lovelace Institute.

Sophie Stalla-Bourdillon is a Professor in Information Technology Law and Data Governance within Southampton Law School and a Senior Privacy Counsel & Legal Engineer at Immuta. Sophie has written extensively on data protection, privacy and governance, and has led the legal effort of various interdisciplinary research projects in the field including EU FP7, Horizon 2020, EPSRC projects. She is Editor-in-chief of the Computer Law and Security Review, and has served as a legal and data privacy expert for the European Commission, the Council of Europe, the OSCE, and the OEDC.

Brian Pickering is a Senior Research Fellow in the School of Electronics and Computer Science, University of Southampton, carrying out research into online behaviours and the acceptance of technology. Using mainly qualitative research methods, he investigates trust relationships and online group formation. Further, as part of application and technology evaluation, he focuses on how potential adopters and users create narratives with technology embedded as a predictor of technology acceptance rather than more traditional models in domains from healthcare to cybersecurity. He is also chair of the Faculty Research Ethics committee and a member of the University DPIA panel.



Steve Taylor is a Senior Research Engineer at the University of Southampton. He received his PhD in Computer Science (Artificial Intelligence) from the University of Greenwich in 1997 and has published over 40 academic papers. His research interests include computational governance & regulation and he is currently investigating a risk management approach applied to regulatory compliance.

Laura Carmichael is a Research Fellow at the Interdisciplinary Centre for Law, Internet and Culture (iCLIC), University of Southampton. She has a PhD in Web Science, and has been involved in various interdisciplinary research projects related to data sharing and re-usage, including the EU-funded Data Pitch open innovation programme.

Executive Summary

Our vision is to improve public health and patient care through responsible data access, collaboration and (re-)usage across academia, the public sector and industry by removing barriers and accelerating existing processes whilst maintaining the highest standards for data governance and security. At the centre of this aspiration lies the establishment of a **Social Data Foundation** ("the Data Foundation") – an innovative trustworthy and scalable data-driven health and social care ecosystem overseen by independent data stewards – created to support multi-party data sharing whilst respecting societal values endorsed by the community.

Our motivation

Future sustainability and performance of health and social care will be achieved through progressive digitisation of organisations, business processes, social communities, and individuals. Digitisation will create new ways to deliver public health, clinical diagnostics, self-health management and prevention, and operations management through a rich ecosystem of connected institutions, people, devices, and data. The process of digitisation will be human-centric to ensure positive change through advice on how decision-making, diagnostics, screening and treatments can be augmented by digital technologies in ways that improve the patient journey.

Collaborative sharing and linking of safe, useful data between different stakeholders under secure and rights-respecting conditions will be vital for positive health and social care transformation both for direct care – e.g. for medical diagnoses, treatments – and indirect care – e.g. for evaluating the effectiveness, efficiency and value of health and social care provision and policies. New links between shared data from multiple agencies in the health and social care sector will allow fundamental rethinking of health and social care systems, care delivery models and workforce capabilities, disrupting traditional functional business units within health and social care systems (e.g. "the hospital", "the community trust", "the GP practice", "residential care home"). To achieve this objective, stakeholders must be convinced of the benefits of multi-party data sharing between care settings – including, increasingly, patients at home – and be confident that security, privacy, and ethical behaviour are ensured. These data will include clinical, social, environmental, and operational data (e.g. admissions, capacity) with varying degrees of de-identification from statistical data to de-identified data at the individual level.

Yet more remains to be done to incentivise, accelerate and join up data sharing amongst stakeholders in ways that are socially acceptable, trustworthy, sustainable, and scalable. While multi-party data sharing initiatives in the health and social care domain are not new, in some cases, existing processes remain disjointed, duplicative, and hard to reconcile with a comprehensive risk-based approach to security and privacy. Citizen engagement also tends to be a weak link. What is more, the extraordinary situation of the global COVID-19 pandemic has only heightened this need – especially for localised intelligence and action that can be scaled and generalised nationally.

Our goal is to **foster a trustworthy data sharing institution called the Social Data Foundation dedicated to improving human health, well-being and safety.** The Data Foundation will include the Southampton City Council, the University Hospital Southampton NHS Foundation Trust and the University of Southampton. Flexible membership will allow other organisations to join and the institution to grow.

Enabling greater data-intensive research & innovation for health & social care whilst respecting societal values

To comply with legal, ethical and cyber-security requirements, multi-party data sharing initiatives must have the **ability to proactively identify and manage risks and breaches** related to the deliberate or unintentional leakage or misuse of data across each stage of the data lifecycle. These compliance and risk activities can be difficult – especially given the lines between non-personal and personal data, and between de-identified patient-level data and confidential patient information, are not always clearly drawn.

Multi-party data sharing initiatives must further **recognise the harms (e.g. weak predictive models, cohort bias, non-optimal systems) attributed to lack of data discoverability and partial or complete omission of data sharing** – and therefore must work to enable greater transparency of system transformation, fairness of innovation opportunities (whilst considering the potential for anti-competitive practices) and overcoming of digital divides.

Despite the **fundamental need for a robust compliance and risk management programme** (e.g. intellectual property rights management and clearance, data protection impact assessments, application of the 5 safes framework, implementation of security standards), this alone is not enough to engender trust and confidence. To achieve social acceptance (i.e. a social licence), multi-party data sharing initiatives must continue to uphold and appraise **stakeholder approvals** – in particular they must ensure key decision-making processes are shaped through **proactive citizen participation and engagement activities** that are both meaningful and representative.

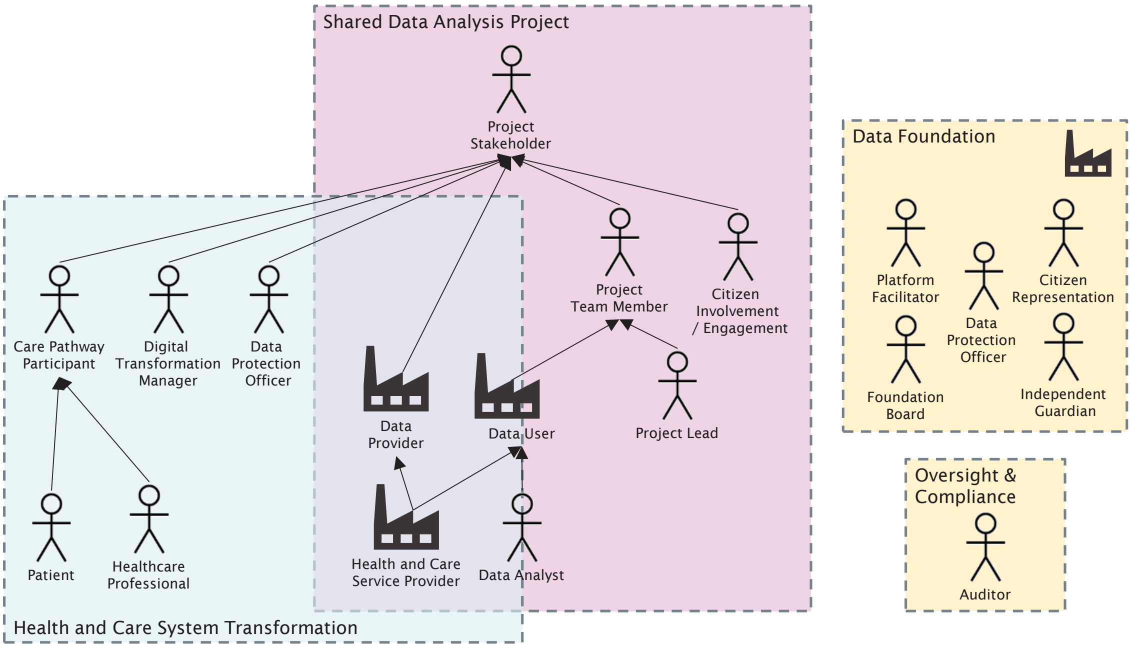


Figure 1. Health and social care transformation

Our approach: the creation of a Social Data Foundation

Striking the balance between the benefits of accelerated health and social care data sharing with risks, as well as sustaining a social licence, is challenging for any multi-party data sharing initiative. Our approach is to create a new data institution called the *Social Data Foundation* that supports multi-party data sharing between the Council, Hospital and University to facilitate health and social system transformation by building on the data foundations framework for good data governance together with strong citizen representation. The key stakeholders and their interactions are shown in Figure 1.

How could this approach work to accelerate existing data sharing workflows?

We believe that the establishment of the Social Data Foundation – as a trusted third-party intermediary (TTPI) – would be able to accelerate existing data sharing as follows:

✓ **BETTER DATA DISCOVERABILITY.** Through a comprehensive metadata catalogue, data users and citizens would have a better understanding about the data available and utility through quality provenance metadata supplied by data providers.

✓ **LOCAL SOLUTIONS WITH NATIONAL LEADERSHIP.** As a localised hub for data-intensive research and innovation and positive health and social care transformation, the Social Data Foundation would be able to promote greater collaboration, address key local priorities and rapidly respond to new and emerging health data-related challenges, whilst offering national exemplars of health system solutions that can be replicated.

✓ **EMPOWERING CITIZENS TO CO-CREATE AND PARTICIPATE IN SYSTEMS TRANSFORMATION.** By widening the range of stakeholders involved in key decision-making processes (data governance, design, evaluation, etc.) to include providers, civil society and communities, and supported through technology-enabled governance, all stakeholders will be better informed about needs and expectations increasing likelihood of data sharing, participation and successful adoption of proposed changes.

✓ **GREATER ASSURANCES THAT BEST PRACTICE DATA GOVERNANCE IS FOLLOWED.** Building trust and confidence between stakeholders in Data Foundation operations is necessary to ensure safe and useful data sharing.

✓ **FASTER ETHICAL OVERSIGHT AND INFORMATION GOVERNANCE.** As a TTPI, the Social Data Foundation would offer semi-automated business processes to rapidly establish approval requests, risk assessment (e.g. de-identification standards) and platform data-flows necessary for institutional and national approval requests (e.g. NHS HRA, NHS REC or Confidentiality Advisory Group (CAG)).

Recommendation: Query-Based Deployment Scenario

The Social Data Foundation can be viewed as a **federation of stakeholders** each with varying degrees of authority and/or influence over decision-making processes. Core data-related functions (e.g. de-identification, data hosting) can be arranged within the federation in numerous ways. Given **the precise distribution of function is directly linked to governance and risk**, and dispersal of control, agency and trust between stakeholders, we have devised four platform deployment scenarios that span the spectrum of federation options from loose to tightly integrated federations of stakeholders:
(i) decentralised, (ii) distributed host,
(iii) centralised and (iv) query-based.

Based on our analysis, we consider the centralised and query-based scenarios to be the most well-suited options for deployment for the following key reasons:

✓ **HIGH PLATFORM COLLABORATION** through collective decision-making body with influence over data discovery, access, and controls.

✓ **HIGH PLATFORM UTILITY** through involvement in core data preparation and data query functions.

- ✓ **HIGH DATA FINDABILITY** through data discovery for all data shared by data providers.
- ✓ **HIGH DATA ACCESSIBILITY** through a single data query made directly for one or more shared datasets held by several data providers.
- ✓ **HIGH DATA ASSURANCE** through data functions executed either in collaboration or independently from data providers with accountable oversight, maximising data stewardship and data sharing advocacy.

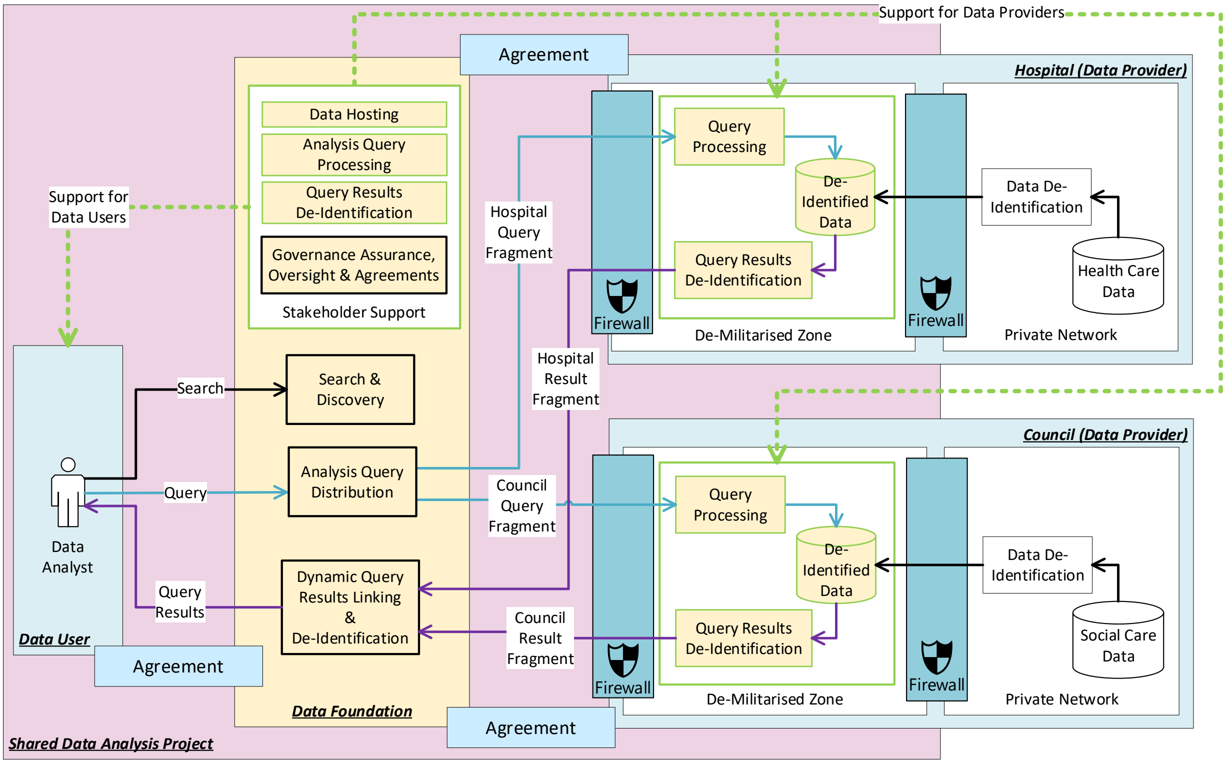


Figure 2. Platform Deployment Scenario 4: Query-based multi-party data sharing initiative

- ✓ **DYNAMIC LINKING** allows for more a granular approach to the utility-privacy trade-off to establish the optimal level of utility for each shared data analysis project whilst preserving privacy through targeted technical and organisational measures.

While we recognise that a centralised approach may be a more realistic option for the immediate operation of a Social Data Foundation, we recommend that work to advance towards a query-based model should begin from the outset. Note that Figure 2 shows the Data Foundation, two data providers and one data user operating the query-based model.

The query-based approach is preferable as there is minimised data replication, retention and associated costs, as data is not stored centrally beyond the needs of specific projects at the point of use, with potential for "caching" and reconstruction.

Key operating principles

The six key operating principles of the Social Data Foundation are as follows:

- **PRINCIPLE 1 – The Social Data Foundation acts as a trusted third party intermediary (TTPI) to facilitate shared data analysis projects** via governance, brokerage of agreements between data providers and data users, shared data management and assurance services, a front-end portal, and tooling to enable sharing operations that are executed at data providers (e.g. for de-identification).

- **PRINCIPLE 2 – The Social Data Foundation provides a dynamic linking service** for authorised data users where two or more sources of health and social care data are brought together on demand according to the specific parameters of an authorised data user's query where the risk of re-identification is both evaluated before and after data linkage, and mitigated through assurance processes facilitated by the Data Foundation.

- **PRINCIPLE 3 – The extent of "data sharing" is limited to the results of pre-approved queries agreed by all parties in a project – not whole datasets.** The Data Foundation facilitates a process to approve queries based on a risk assessment and provides a gateway for data analysis queries from authorised data users.

- **PRINCIPLE 4 – The Social Data Foundation carries out a risk assessment for each shared data analysis project before any data is shared** by data providers and assigns a list of pre-approved queries to authorised data users.

- **PRINCIPLE 5 – Data providers only share de-identified data as part of the Social Data Foundation.** The possible risk of re-identification – related to a specific pre-approved data analysis query – is addressed at the point of delivery by each data provider before their data is linked with other data providers' data, as well as at the point of linking at the Data Foundation and mitigated through assurance processes facilitated by the Data Foundation.

- **PRINCIPLE 6 – Agreements govern relationships between all stakeholders** for each shared data analysis project, including the assignment of pre-approved queries to one or more authorised data users as part of a specific project.

1. Introduction

Our vision is to improve public health and patient care through responsible data access, collaboration and (re-)usage across academia, the public sector and industry by removing barriers and accelerating existing processes whilst maintaining the highest standards for data governance and security. At the centre of this aspiration lies the establishment of a **Social Data Foundation** ("the Data Foundation") – an innovative trustworthy and scalable data-driven health and social care ecosystem overseen by independent data stewards – created to support multi-party data sharing whilst respecting societal values endorsed by the community.

Our motivation

Future sustainability and performance of health and social care will be achieved through progressive digitisation of organisations, business processes, social communities, and individuals.[1] Digitisation will create new ways to deliver public health, clinical diagnostics, self-health management and prevention, and operations management through a rich ecosystem of connected institutions, people, devices, and data. The process of digitisation will be human-centric to ensure positive change through advice on how decision-making, diagnostics, screening and treatments can be augmented by digital technologies in ways that improve the patient journey.[2]

Collaborative sharing and linking of safe, useful data[3] between different stakeholders under secure and rights-respecting conditions will be vital for positive health and social care transformation both for direct care – e.g. for medical diagnoses, treatments – and indirect care – e.g. for evaluating the effectiveness, efficiency and value of health and social care provision and policies.[4] New links between shared data from multiple agencies in the health and social care sector will allow fundamental rethinking of health and social care systems, care delivery models and workforce capabilities, disrupting traditional functional business units within health and social care systems (e.g. "the hospital", "the community trust", "the GP practice", "residential care home"). To achieve this objective, stakeholders must be convinced of the benefits of multi-party data sharing between care settings – including, increasingly, patients at home – and be confident that security, privacy, and ethical behaviour are ensured. These data will include clinical, social, environmental, and operational data (e.g. admissions, capacity) with varying degrees of de-identification from statistical data to de-identified data[5] at the individual-level.[6]

Yet more remains to be done to incentivise, accelerate and join up data sharing amongst stakeholders in ways that are socially acceptable, trustworthy, sustainable, and

[1] For the purposes of this white paper, we define the term 'progressive digitalisation' as follows: the transformation of large and complex systems from analogue to digital form through incremental steps in different parts of the overall system.

[2] As a benchmark for best practice, Scott et al. (2018) outline the following five-point framework for evaluating whether a potential data sharing activity can be considered to be of public benefit (for full information, including the scale and questions, see original document): *"1. That it enables high quality service delivery which produces better outcomes for people, enhancing their wellbeing; [/] 2. That it delivers positive outcomes for the wider public, not just individuals; [/] 3. That it uses data in ways that respect the individual, and their privacy, not just in the method of sharing but also in principle; [/] "4. That it both represents and supports the effective use of public resources (money, time, staff) to enable the delivery of what people need/want from public services; [/] "5. That the benefits are tangible, recognised and valued by service providers and the wider public"* (Scott et al., 2018).

[3] This phrase reflects the terminology used by the UK Anonymisation Network (UKAN) Decision-Making Framework (Elliot et al., 2016). In other words, health and social care data that are *"purposeful"*, *"proportionate"* and *"responsible"* (Scott et al., 2018).

[4] For more information about data sharing for direct and indirect care purposes refer to the Information Governance Review (2013).

[5] For the purposes of this white paper, we define 'de-identified data' as follows: Individual-level data that has been subject to both data and process controls such that the re-identification risk can be considered to be remote. De-identified data should be considered to meet the legal standard for anonymisation.

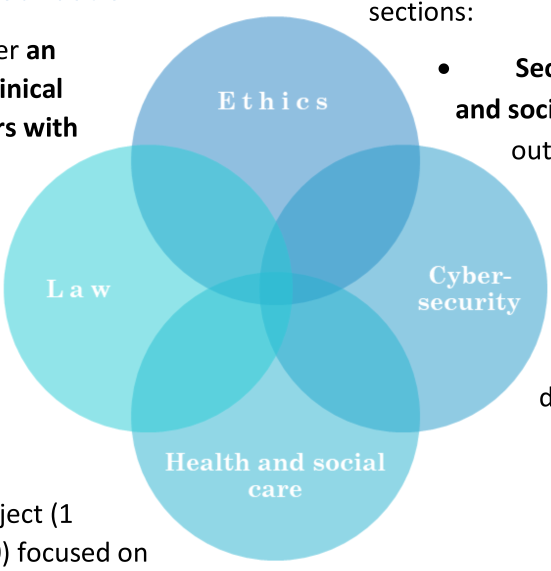
[6] An example of key-patient level data attributes required for health care simulation modelling would include (CORMSIS, 2020): age of patient; gender identity of patient; diagnostic codes (ICD106 or similar) on admission and on discharge of patient; time-stamped data showing admission of patient; transfer(s) of patient from one unit to another within the hospital; length and type of procedure(s) (where applicable); discharge destination for patient (typically recorded as the patient's usual place of residence; however it is also often useful to have clinical context such as type of residence e.g. own home, nursing home); and, the provider and information for clinical outcomes related to the patient.

scalable. While multi-party data sharing initiatives in the health and social care domain are not new,[7] in some cases, existing processes remain disjointed, duplicative, and hard to reconcile with a comprehensive risk-based approach to security and privacy. Citizen engagement also tends to be a weak link. What is more, the extraordinary situation of the global COVID-19 pandemic has only heightened this need – especially for localised intelligence and action that can be scaled and generalised nationally.[8]

Our goal is to foster a trustworthy data sharing institution called the Social Data Foundation dedicated to improving human health, well-being and safety. The Data Foundation will include the Southampton City Council, the University Hospital Southampton NHS Foundation Trust and the University of Southampton. Flexible membership will allow other organisations to join and the institution to grow.

About the Social Data Foundation

The initiative brings together **an interdisciplinary team of clinical and social care practitioners with data governance, health data science, and security experts** from ethics, law, technology and innovation, web science and digital health in order to design and develop a Social Data Foundation ("the Data Foundation"). An initial project (1 June to 30 September 2020) focused on accelerating responsible access, collaboration and (re-)usage of health and social care data between three partners: (1) Southampton City Council ("the Council"), (2) the University Hospital Southampton NHS Foundation Trust ("the Hospital") and (3) the University of Southampton ("the University").



The Social Data Foundation Project is partly funded and supported by the University of Southampton's Web Science Institute (WSI) and Southampton Connect. Southampton Connect is an independent partnership of senior city representatives drawn from a wide-range of sectors, including the Solent NHS Trust, Southampton City Council and the University.

White paper: objective and overview

The objective of this white paper is to provide our rationale as well as outline an initial architecture specification for the creation of new data institution[9] – i.e. a Social Data Foundation.

This white paper is divided into three main sections:

- **Section 2. Facilitating health and social care transformation** – outlines the wider health and social care strategy for Southampton and surrounding areas, and further provides four exemplary use cases to demonstrate the need for a

[7] For further background, Jones & Ford (2018) provide a basic dichotomy of data sharing models; also see Appendix A to this white paper for a list of examples.

[8] Note there are numerous barriers to health and social care data sharing – van Panhuis et al. (2014) outline six main categories of barriers related to public health data sharing for public health: (i) *"Technical barriers" – e.g. "Lack of metadata and standards"; (ii) "Motivational barriers" – e.g. "Disagreement on data use"; (iii) "Economic barriers" – e.g. "Lack of resources"; (iv) "Political barriers" – e.g. "Restrictive policies"; (v) "Legal barriers" – e.g. "Protection of privacy"; and (vi) "Ethical barriers" – e.g. "Lack of proportionality"* (van Panhuis et al., 2014). For further background information, Sane & Edelstein (2015) examine the ways in which these six types of barriers may be overcome.

[9] Note that the phrase 'data institution' is used by the Open Data Institute (ODI) as an umbrella term to describe *"organisations whose purpose involves stewarding data on behalf of others, often towards public, educational or charitable aims"* (Dodds et al., 2020), such as data trusts (Hardinges & Tennison, 2020).

trusted third-party intermediary (TPPI)[10] for data governance.

- **Section 3. Key governance requirements for trusted third party intermediaries (TTPI)** – focuses on cyber security and data protection risk management, the need for robust citizen representation and provides a non-exhaustive list of some key data governance requirements to be fulfilled by a TTPI.
- **Section 4**. **Our approach: the creation of a Social Data Foundation** – provides an overview of the data foundation framework, collaborative data scenarios, a proposed data governance structure, and examines four platform deployment scenarios for the Data Foundation.

We conclude this white paper (section 5) by summarising our findings from sections 2, 3, and 4 – including our recommended platform deployment.

[10] For the purposes of this white paper, we define the term 'trusted third party intermediary' (TTPI) as follows: a responsible and reliable entity that facilitates data sharing interactions for projects related to health and social care transformation between data users and data providers, whose involvement is acceptable to all parties involved.

2. Facilitating health and social care transformation

A primary requirement for establishing a trustworthy data sharing alliance between the Council, Hospital, and University is to facilitate a wide-range of use cases related to sustainable and positive health and social care transformation across academia, the public sector and private sector. This section therefore outlines the wider health and social care strategy for Southampton, and further provides four exemplary use cases to demonstrate the need of a TPPI for data governance.

An overview of integrated care systems (ICS)

In 2020, Hampshire and Isle of Wight (HIOW) established a shadow Integrated Care System (ICS) with formal status from April 2021.[11] The ICS creates closer collaboration between all NHS bodies (primary care, acute care, community services, mental health services, etc.) in the region serving the 1.8 million population. The Southampton City 5-year health and care strategy outlines a vision which fits within the broader Integrated Care Partnership for Southampton and South West Hampshire. The University Hospital Southampton NHS Foundation Trust Hospital is positioned within the wider health system as a large general hospital with a diversity of acute care services that is unique outside London and other larger metropolitan areas. In addition, the Hospital is one of the biggest providers of specialised services in England serving a 3.7m population as a specialist centre. The geographic region and environmental conditions are highly diverse and include urban, maritime, and rural economic activities, and large permanent and transient populations presenting a wide range of health and care needs. The health and care services and population factors make the region highly attractive for discovery and evaluation of new data-driven healthcare technologies.

ICSs bring together services, processes, data, and people responsible for disease management, mental health, and community care with socio-economic needs for education, mobility, housing, environment, social capital, and financial support. Failure to address social requirements can lead to poor self-management and medication adherence, health inequalities, social isolation, financial problems, and deteriorating health. Integrated care is expected to enable the shift from disease-centred to holistic person-centred care and offer greater opportunities for personalised medicine and care (Kyriazis et al., 2017).

Four use case exemplars

We have worked with stakeholders to elaborate four use cases as exemplars to show how trustworthy data sharing can lead to data-driven solutions targeting better care. These use cases, outlined below, are not exhaustive but nonetheless demonstrate the need for governance facilitated by a TTPI.

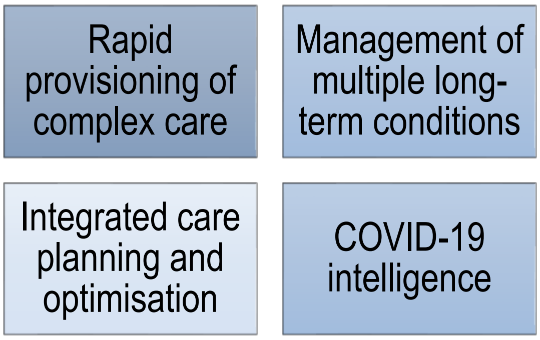

Rapid provisioning of complex care
Management of multiple long-term conditions
Integrated care planning and optimisation
COVID-19 intelligence

Figure 3. Four use case exemplars

[11] In 2016, NHS organisations and local councils came together to form sustainability and transformation partnerships (STPs) covering the whole of England, and set out their proposals to improve health and care for patients. In some areas, a partnership will evolve to form an integrated care system, a new type of even closer collaboration.

(1) Rapid provisioning of complex care

(a) Better care need:

Today, it is difficult to find the right care packages to meet specific patient needs, whether this be complex care resources to recover from hospital, an available residential place, a support group, or healthy physical activity. Given the huge range of public and private providers, the complexity of individual and family care requirements, the lack of curated access to market information, and the lack of guided support in choosing the most appropriate care option, solutions are needed for a social care service marketplace to support rapid discovery and access to services.

Data has the potential to allow the development of new and novel deliveries of information, feed into smart city planning and intelligent commissioning, harness benefits of AI and machine learning, improve family and ageing outcomes.

(b) Governance needs:

Care marketplaces should establish a statutory family information ecosystem through a multi-agency infrastructure that:

- Makes it possible to join up these legacy platforms;
- Provides novel search functions for care services;
- Enables local authorities to better meet information statutory duties; and
- Empowers families to self-serve and make informed choices to meet their specific and unique needs.

Medical records, care plans, care service directory and potential for new open datasets are all related to care and family services. Often several decision-makers must collaboratively work together to plan care and allocate resources.

A TTPI would provide a place where stakeholders could explore re-configuration and re-design of pathways and decision processes at a system level through data sharing.

(2) Management of multiple long-term conditions

(a) Better care need:

A substantial number of people (30% all ages, 54%>65 and 83%>85) suffer from two or more long-term conditions, sometimes referred to as multi-morbidity (Cassell et al., 2018). Having multiple long-term conditions affects quality of life, leads to poorer health outcomes and experiences of care, and accounts for disproportionate healthcare workload and costs. Solutions are needed to understand disease trajectories over the life-course (start well, live well, age well) at population levels, and to develop effective personalised interventions.

(b) Governance needs:

Complex and heterogeneous longitudinal and real-world linked data is needed to study the clusters and trajectory of disease.

A TTPI would support new data flows to accelerate data sharing and learning using query-based linking to reduce replication and retention, whilst federated learning would support population level learning without moving data. Regionally shared care record initiatives (such as the Wessex Care Record, HIOW Public Health Management and HIOW Care and Health Information Exchange (CHIE)/Care and Health Information Analytics (CHIA)) publish guidelines for Information Governance (IG) building on national and regional activities (such as the HIOW Sustainability and Transformation Partnership (STP) IG working group).

(3) Integrated care planning and optimisation

(a) Better care need:

Health and care systems have limited resources, and providers need to plan capacity and optimally allocate resources, often in real-time, to achieve sustainability and performance targets (patient throughput, waiting times, staff utilisation, patient satisfaction).

The complexity of the NHS makes it difficult to track individual patients between care settings, and links with non-NHS organisations increase this complexity further. The challenge for integrated care is that demand and service provisioning is distributed between multiple providers. Understanding and predicting the utility of resources and side effects of system commissioning, pathway design choices and provider decisions requires new operational approaches for planning, optimisation, and collaborative decision-making.

(b) Governance needs:

Exploring operational management of integrated care systems has a strong requirement for linked data across organisations. Data should include medical records for variability in patient characteristics, realistic service demand patterns, service usage, and resource usage.

A TTPI can support collaborative multi-stakeholder working required to explore integrated pathways, commissioning models and the impact of local and global decision making.

A TTPI would also allow for new governance models to be explored that consider emerging city infrastructure and institutional operating models underpinning initiatives, such as Digital Twins.

(4) COVID-19 intelligence

(a) Better care need:

Intelligence about prevalence of infection and the number of susceptible individuals in populations are critical to public health policies, health system capacity planning and social distancing policies in response to pandemics such as COVID-19. There is a need for population sampling and symptom reporting within NHS Test and Trace service, where data are aggregated to provide clusters and population level statistics as input to intelligence and outbreak control.

There is a need to bring together social network models, individual behavioural models, epidemiological models, and clinical models into a coherent situational assessment framework.

(b) Governance needs:

Population-level surveillance of a single parameter of health is relatively simple considering the relationship between data providers and citizens. Data providers (General Practice, Southampton City Council, University of Southampton) provide access to the population samples whilst a further data provider (the Hospital) holds the medical record and test results. The data is limited and the purpose of data usage and analysis well defined, being one source of testing data (alongside other pillars of testing) provided to the NHS Test and Trace service.

A TTPI would allow for extended modelling scenarios including forecasting infection in populations or capacity planning by linking datasets such as mobility, social networks/interactions, and health data (infections, deaths, hospitalisations). The data providers are diverse and have different motivations, which raises governance challenges such as the secondary use of datasets.

3. Key governance requirements for trusted third party intermediaries (TTPI)

This section describes some of the critical data governance requirements to be fulfilled by a TTPI – to ensure trustworthy, secure and safe data sharing that respects the rights and freedoms of data subjects, and unlocks the benefits of multi-party data sharing for health and social care transformation.

Cyber security and personal data protection risk management

To comply with data governance requirements, multi-party data sharing initiatives must proactively identify and manage risks and breaches related to the deliberate or unintentional leakage or misuse of data across each stage of the data lifecycle.[12] A recent report published by the OECD (2019) provides the following four key categories of risks and challenges related to enhanced data sharing, usage and re-usage:

- *"Digital security risks and confidentiality breaches […]"*;
- *"The violation of privacy, intellectual property rights and other interests"*;
- *"The difficulty of applying a risk management approach", including "Challenges of managing the risks to third parties"*; and
- *"Barriers to cross-border data access and sharing […]"* (OECD, 2019).

These compliance and risk activities can be difficult in practice – especially without effective monitoring and oversight of (re-)usage practices, and that the lines between non-personal and personal data, and between de-identified patient-level data and confidential patient information, are not always clearly drawn.

Multi-party data sharing initiatives must further **recognise the harms (e.g. weak predictive models, cohort bias, non-optimal systems) attributed to lack of data discoverability and partial or complete omission of data sharing** (Jones et al., 2017) – and therefore work to enable better transparency of system transformation, fairness of innovation opportunities (whilst considering the potential for anti-competitive practices) and help overcome digital divides.

Cyber security certification and risk assessment

Increasingly, certification using standardised information security assurance processes is required to provide evidence and assurance that an organisation is competent regarding security resilience and risk assessment. For example, powerful partners (e.g. blue-chip companies) are demanding ISO 27001 certification[13] of their supply chain partners.[14] Furthermore, Cyber Essentials is a UK government-backed scheme to protect UK businesses from cyber-attacks by the assessment of threats and implementation of basic mitigation measures. Compliance can either be self-assessed or certified by an outside body ("Cyber Essentials Plus"). For

[12] For instance, Article 11 of the OECD Recommendation of the Council of Health Governance recommends that organisations should implement controls and safeguards that include: *"formal risk management processes, updated periodically that assess and treat risks, including unwanted data erasure, re-identification, breaches or other misuses, in particular when establishing new programmes or introducing novel practices"* (OECD, 2016).

[13] ISO 27001 involves identification of risks and appropriate risk management mechanisms (covered by ISO 27001 and related standards which build on a more generic risk management approach defined by ISO 31000 and ISO 31010); and verification of the security properties in the system implementation (covered by ISO 15408). This provides a standardised mechanism to check whether identified threats are addressed by determining security risks and specifying measures that (if correctly implemented) will address those risks.

[14] Note that other key ISO standards for information security include ISO 27005, ISO 27002 and other ISO 27000-series standards. ISO 27005 approaches risk assessment by considering an information system as a set of assets that should be protected from threats that may compromise its function. Security controls can then be introduced within the system to prevent or mitigate the threats. ISO 27005, ISO 27002 and other ISO 27000-series standards provide check lists for some types of threats and security controls. These controls are also not limited to technical IT mechanisms, and may include, e.g. legal, organisational or physical controls.

Health Data, the NHS Digital Toolkit requires annual self-assessment regarding information security and personal data protection as a pre-requisite to access NHS data.

Certified compliance provides assurance and confidence for potential and existing partners that an organisation understands cyber security threats and data protection, and has installed measures to mitigate risks and protect personal data. A TTPI therefore needs to be certified to be credible.

Key misbehaviours, threats and mitigating controls related to data sharing

ISO 27005 describes a risk management approach that concerns *assets,* which may be compromised by *threats.* If a threat compromises an asset, then the asset may *misbehave*. The *risk* to the asset is the combination of the likelihood of the threat occurring with the severity of the misbehaviour of the asset to its owner or other stakeholders. *Controls* are applied to the asset to mitigate the risks to it by either lowering the likelihood of the threat occurring or reducing the severity of the impact to the asset should the threat occur.

Data are the key assets of concern for the TTPI – the protection of these data is therefore its highest priority. This does not only relate to the data managed by the TTPI, but also needs to ensure that the actions and purposes of the TTPI cannot compromise any data within the data providers. From a cybersecurity perspective, the key misbehaviours for data, common threats that cause them and controls that mitigate the risks are listed in Table 1.

The TTPI therefore must conduct valid risk assessments and implement suitable controls to protect the confidentiality, integrity, and availability of the data it holds or manages. Of note, risk assessment will go beyond the traditional cybersecurity triad (availability, integrity, confidentiality) to meet key data protection goals such as data minimisation, storage limitation, purpose limitation, data accuracy, and lawfulness, fairness and transparency.

Table 1. Overview of key misbehaviours, threats and controls related to data sharing

MISBEHAVIOUR	THREATS	CONTROLS
Loss of Availability - data is not available to authorised actors when it is needed	Server compromise or failure	Maintenance and patching of host server
	Data inaccessible to authorised users	Access control for data Key management of data Maintenance and patching of host server
Loss of Integrity - data is compromised by e.g. corruption or unauthorised alteration. Can cause *Loss of Authenticity* - data is no longer considered reliable (authentic)	Corruption of stored data by compromised host	Encryption of data at rest and in transit
	Corruption of data	Access control for data Encryption of data at rest and in transit
Loss of Confidentiality - unauthorised actors access data	Leaking of data due to compromised host	Encryption of data at rest Maintenance and patching of host server
	Data accessible to unauthorised users	Encryption of data at rest Maintenance and patching of host server
	Gap in end-to-end encryption of data	Access control for data Encryption of data in transit

Managing re-identification risks

As part of health and social care transformation, data users may (re)use a single-source of data, or multiple sources of data – referred to as “data linking”. The Public Health Research Data Forum (2015) define the term “data linking” as follows: *“bringing together two or more sources of information* *which relate to the same individual, event, institution or place. By combining the information it may be possible to identify relationships between factors which are not evident from the single sources".*

While data linkage may have various benefits for a wide-range of stakeholders,[15] it may also increase the risk of re-identification of data subjects (e.g. patients, service-users). **One of the key risks that must be managed by a TTPI therefore is the potential for re-identification of data subjects that can arise through data sharing, usage and re-usage.**[16] Oswald (2013) defines the risk of re-identification as: *"the likelihood of someone being able to re-identify an individual, and the harm or impact if that re-identification occurred."*[17]

In addition to "linkability", there are two other key ways in which individuals may be re-identified: by "singling out" individuals, and via "inference" – i.e. deducing some information about an individual (Article 29 Data Protection Working Party, 2014). As part of a robust risk management programme, it is therefore essential that a TTPI can *"mitigate the risk of identification until it is remote"* (Information Commissioner's Office, 2012), and ensure such risks are periodically reviewed.

The TTPI must take a data protection by design and by default approach (as per Article 25 of the GDPR) by identifying and implementing appropriate organisational and technical safeguards, e.g. put in place prevention measures to mitigate re-identification risks; in particular, the TTPI should practise "functional anonymisation" (Elliot et al., 2018). The phrase 'functional anonymisation' is defined by Elliot et al. (2018) as: *"the practice of reducing the risk of re-identification through controls on the data and its environment so that it is at an acceptably low level"*.

Preserving the utility of data

There is a suggestion that greater protection of personal data leads to safer *but* less useful data – often referred to as "the privacy-utility trade-off". However, this is not always the case – for instance, in circumstances where data subjects place more trust and confidence in data protection safeguards, the utility of data can also increase as *"it may lead to greater willingness of data subjects to provide accurate data in the first place"* (Elliot et al., 2018). **The TTPI must be able to protect personal data effectively and appropriately as well as preserve the utility of data for the purposes of positive health and social care transformation.**[18]

Proactive citizen representation and engagement

Despite the fundamental need for a robust compliance and risk management programme – e.g. intellectual property rights management and clearance, data protection impact assessments (DPIAs), application of the 5 safes framework, implementation of security standards – this is not enough alone to engender trust and confidence. To achieve social acceptance (i.e. a social licence (Carter et al., 2015; Jones & Ford, 2018)), multi-party

[15] Examples of data linkage benefits include *"Increasing the range of feasible topic areas" such as "identifying the correlation between health events from different sources" and "identifying contributory factors from non-health data"; "generating useful tools"; "improved use of scarce resources", such as for "international comparison" and "interdisciplinary research"* (Public Health Research Forum, 2015).

[16] It is worthwhile to note that pursuant to s.171(1) of the Data Protection Act 2018 that: *"It is an offence for a person knowingly or recklessly to re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data."*

[17] For further guidance on the risk of re-identification see: Information Commissioner's Office (2012). Anonymisation: managing data protection risk code of practice (2012); for further background information also refer to: Alexin (2014).

[18] For instance, Recital 4 of the GDPR states: *"The processing of personal data should be designed to serve mankind. The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality. […]"*The ICO (2012) also acknowledges *"the special utility of personal data and that it is not always necessary or possible to use anonymised data instead of personal data"*.

data sharing initiatives must continue to uphold and appraise stakeholder approvals – in particular they must ensure key decision-making processes are shaped through proactive citizen participation and engagement activities that are both meaningful and representative.

While such citizen engagement and participation is not new, it is in receipt of close attention,[19] especially as health and social care research and innovation becomes increasingly data-driven (Aitken, 2019) and disconnected from data subjects, e.g. with increased secondary use of health and social care data (Jones & Ford, 2018)[20] where such re-use is often less understood (CurvedThinking, 2019).

A key area of focus for the health and social care domain is how to improve citizen engagement and participation[21] in decision-making processes.[22] There are three main categories of citizen engagement and participation, outlined by Aitken et al. (2019), as follows:

- *"Awareness-raising"* activities – e.g. *"training", "media campaigns", "website";*
- *"Consultation"* activities – e.g. *"interviews", "focus groups", "surveys"* ; and
- *"Empowerment"* activities – e.g. *"advocacy", "participatory appraisal", "workshops"* (Aitken et al., 2019).

Furthermore, in some cases citizen engagement and participation activities may fall across two or all these categories, e.g. *"citizen juries"* (Aitken et al., 2019). Many data-intensive data sharing initiatives also have consumer panels (Jones & Ford, 2018). For instance, the Public Benefit and Privacy Panel for Health and Social Care (NHS Scotland) has been set up by NHS Scotland as a data advocacy panel to scrutinise access to health data for non-direct care.[23]

As a method of public engagement, feedback to citizens on the benefits that have arisen from insights derived from linked data may be appropriate – which will contribute to overall transparency.

Summary of data governance requirements

While not an exhaustive list, we provide some key data governance requirements to be fulfilled by a TTPI:

KEY DATA GOVERNANCE REQUIREMENTS

1. Certification

a. The TTPI must have certified compliance with UK Cyber Essentials Plus.

[19] For instance, the Medicines and Healthcare Products Regulatory Agency (2020), the UK regulator for medicines, medical devices and blood components for transfusion, is currently improving its mechanisms for citizen participation based on feedback from a 12-week public consultation.

[20] Sets of guiding principles for citizen participation in health and social care are emerging, including: the consensus statement for public involvement and engagement in data intensive health research (Aitken, 2019); National Standards for Public Involvement in Research (2019); and ten principles form Health Data Research UK (2020a).

[21] Throughout this white paper, we use the term 'citizen' in connection with participation and engagement akin to the NHS terms *"patient and public participation"* and *"patient and public engagement"* to emphasise the inclusion of social care interaction as well as healthcare.

[22] E.g., at a national level, the NHS Citizen programme 2013-2016 (Involve, 2020a; The National Archives, 2020), NHS England Patient and Public Participation Policy (Public Participation Team, 2017). There are also twelve clinical networks and twelve clinical senates across England, including the Wessex Clinical Network and Wessex Clinical Senate that cover Southampton, Hampshire and the Isle of Wight (NHS England, 2020). Also note that Involve (2020b) provides public engagement training.

[23] The SAIL Databank provides another useful example of an existing programme for citizen participation within the area of health and social care. This programme is set out through its Public Involvement & Engagement Policy (Jones, 2020). In particular, the SAIL Databank Ladder of Public Involvement & Engagement Activities (Jones, 2020) – adapted from Arnstein (1969) – provides a spectrum of citizen participation: *"Community-owned initiatives – are led from start to finish by the public, including decision-making [/;] Co-production – is enabling people to be equal partners in developing and making decisions on an endeavour [/;] Inclusion – is actively involving and/or engaging people in shaping an endeavour [/;] Consultation – involves asking people their views, sometimes in a formal process, with opportunity for input of views [/;] Tokenism – is including a person(s) to tick the box, with little use of their views [/;] Informing – is about providing people with information on a oneway basis [/;] Therapy – is designed to tell people what to think or to dispel views held"* (Jones, 2020).

b. The TTPI must perform annual self-assessment compliant with NHS Digital Toolkit (if NHS data is required).

c. The TTPI should have certified compliance with ISO 27001 risk management.

2. De-identification

a. The TTPI must ensure that all data to be provided are de-identified to acceptable standards (i.e. meets the legal standard for anonymisation[24]).

b. The TTPI must monitor all data linkage queries to assess the risk of re-identification from their results in order to prevent the release of linked data that are not de-identified to acceptable standards.

c. The TTPI must take a data protection by design and by default approach by identifying and implementing appropriate organisational and technical safeguards, e.g. put in place prevention measures to mitigate re-identification risks; in particular, the TTPI should practise “functional anonymisation” (Elliot et al., 2018).

3. Privacy (other than confidentiality)

a. The TTPI must ensure that accurate data is processed for one or more specified and legitimate purposes, and the quantity of the data is limited to what is necessary for these purposes. Each data sharing project should be based upon a valid legal basis and processing within these projects should be fair. Each project should be auditable.

b. The TTPI must monitor implementation of data policies per project and per purpose.

c. The TTPI must take a data protection by design and by default approach by identifying and implementing appropriate organisational and technical safeguards, e.g. implementing purpose-based access control and data policies per project and per purpose.

4. Ethics approvals and permissions[25]

a. The TTPI must attain all necessary ethics approvals and permissions (as applicable) and ensure that they are in place before any data are processed (e.g. Integrated Research Application System (IRAS), Ethics and Research Governance Online (ERGO) 2).

b. The TTPI must ensure that any amendments to data processing activities receive further ethical approval (as applicable).

c. The TTPI must ensure that end of research and/or innovation activities declarations are made to ethics bodies and other organisations (as applicable).

d. The TTPI must determine an explicit set of ethical requirements that clearly describe the core principles and expected standards of behaviour for all stakeholders to uphold (e.g., the Caldicott principles (Department of Health, the Caldicott Committee, 1997; The UK Caldicott Guardian Council), the CARE Principles for Indigenous Data Governance (2018)[26]).

5. Contractual obligations

a. The TTPI must comply with its contractual obligations, e.g. to sub-license data to data users from data providers.

b. The TTPI should choose an appropriate licence to encourage sharing whilst maintaining provenance information.

6. Citizen participation

a. The governing body of the TTPI must devise a strategy for meaningful, proactive citizen participation and ensure that it is implemented and maintained by the TTPI and other stakeholders.

[24] The definition of anonymised data is provided by GDPR Recital 26, namely *“information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.”* Although strictly speaking, Recital 26 is not binding it has been used by the Court of Justice of the European Union and other national courts to interpret the concept of anonymised data. As a matter of principle, two different processes can lead to anonymised data: a risk-based approach to aggregation (i.e., data is aggregated, e.g. to produce counts, average, sums) or a risk-based approach to de-identification (i.e., data remains at the individual level). In both cases, data and context controls should be combined to guarantee that re-identification risk is remote over time.

[25] See Appendix B to this white paper for an ethics checklist for the Data Foundation.

[26] I.e. *“collective benefit”*, *“authority to control”*, *“responsibility”* and *“ethics”* (CARE Principles, 2018).

7. Authoritative guidance

a. The TTPI must follow authoritative best practice for trusted research environments (TREs)[27] and data safe havens.[28] This includes the application of the 5 safes framework – (1) *"safe people"*, (2) *"safe projects"*, (3) *"safe settings"*, (4) *"safe outputs"* and (5) *"safe data"* (UK Data Services) – plus one (6) *"safe return"* (UK Health Data Research Alliance, 2020).

[27] The recently-established UK Health Data Research Alliance (2020) is focused on various areas to improve the re-use of health care data – one such area is on trusted research environments.

[28] For instance, the Information Governance Review (2013) includes thirteen required standards of data stewardship good practice for accredited data safe havens. Furthermore, the Scottish Government (2015) has published a Charter for Data Safe Havens in Scotland for handling unconsented data from NHS patient records to support research and statistics.

4. Our approach: the creation of a Social Data Foundation

Striking the most appropriate balance between the benefits that arise from accelerated health and social care data sharing with its associated risks, as well as sustaining a social licence, is particularly challenging for any multi-party data sharing initiative.

Our approach is to create a new data institution for multi-party data sharing between the Council, Hospital, University called the *Social Data Foundation* – that builds on the data foundations framework (Stalla-Bourdillon et al., 2019; 2020) for good data governance (outlined below) together with strong citizen representation:

SIX FUNDAMENTAL COMPONENTS FOR ANY DATA GOVERNANCE MODEL

(1) *"**A comprehensive rulebook** [/] A data governance model must have a comprehensive rulebook for the usage, sharing, and re-usage of personal and non-personal data, including a robust ethical framework, which should be made publicly available. […]"*

(2) *"**An independent governance body** [/] A data governance model must have a strong, independent governance body comprised of independent data stewards with interdisciplinary expertise. The data steward role should be at the core of any data governance model and oversee the decision-making body. […]"*

(3) *"**An inclusive decision-making body** [/] A data governance model must have a decision-making body that engages participants (in particular data providers) and represents the interests of data subjects. […]"*

(4) *"**A standardised process for flexible membership** [/] A data governance model must have a standardised process to enable relative flexibility in its membership so that: [/] The structure can smoothly grow over time without exaggerated (legal) costs. [/] The risks of harm arising through anti-competitive practices are mitigated, e.g. organisations are not excluded from joining a data governance model without reasonable justification."*

(5) *"**A trust-enhancing technical and organisational infrastructure** […] [e.g.] reduce unnecessary data movements […] Monitor queries. […]"*

(6) *"**A well-regulated legal structure** […] [e.g.] Represent all stakeholders in decision-making processes, e.g. to give data providers – from start-ups to multinational companies – and data subjects rights and opportunities to voice their opinions regardless of their nature, size, or number. […]"*

Source: Stalla-Bourdillon et al., 2019.

To further reduce the re-identification risk associated with data linkage as well as preserve data utility, the Data Foundation would act as a TTPI to offer a safe, secure, and ethical service for dynamic linking for data providers and data users.

Collaboration scenarios

The primary purpose of the Social Data Foundation is to act as a facilitator for health and social care transformation, which is shown by Figure 24. The diagram is organised into three main sections:

- **Blue Box: 'Health and Care System Transformation'** – shows actors working in health and social care who are interested in its transformation.

- **Lilac Box: 'A Shared Data Analysis Project'** – represents the collaboration between different institutions to share data for the purposes of health and social care transformation.

- **Orange Boxes: 'A Data Foundation' with associated 'Oversight & Compliance'** – whose purpose is to facilitate and support the Shared Data Analysis Project.

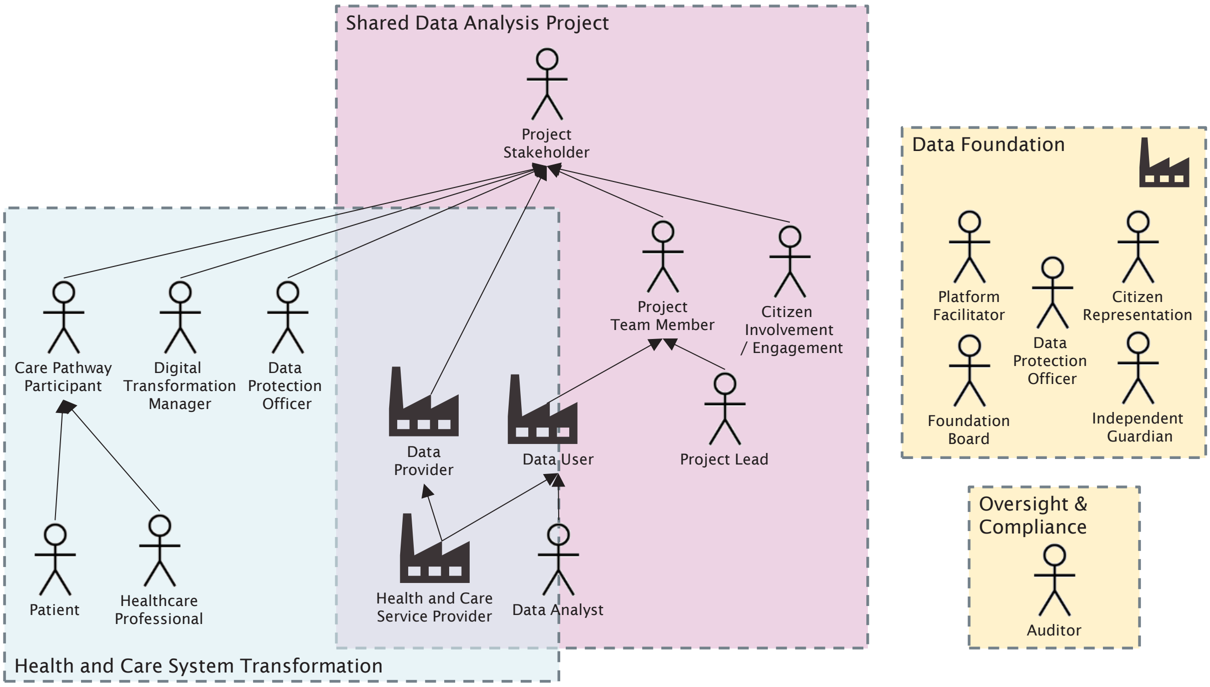


Figure 4. Health and social care transformation

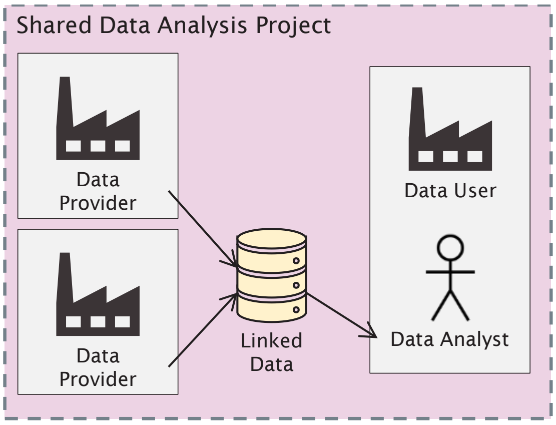


Figure 5. Independent data providers and data users

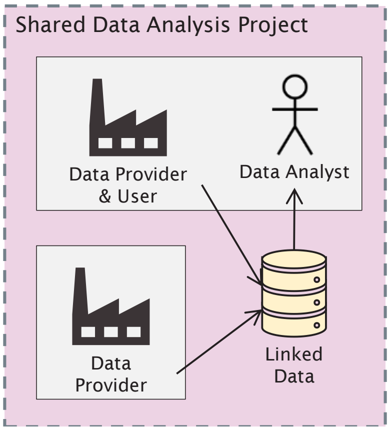


Figure 6. One collaborating organisation is both a data provider and a data user

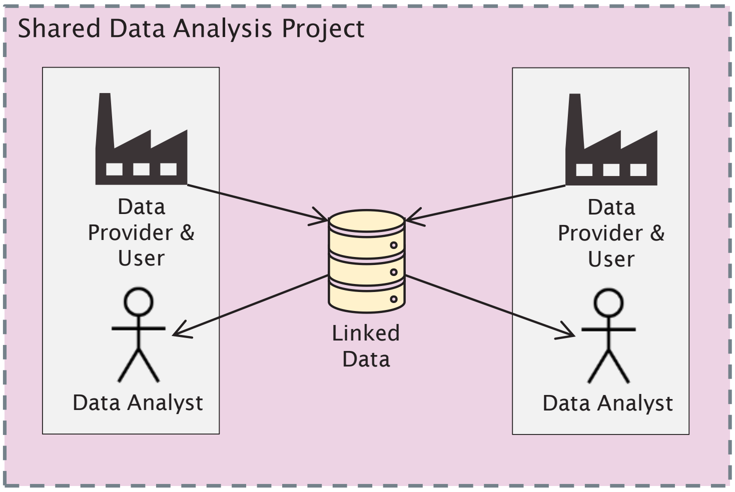


Figure 7. All collaborating organisations are both data providers and users

Figure 5 shows two data providers and an independent data user, meaning that the data analyst is not a member of either of the data provider institutions. While this situation is possible, strong safeguards for security, privacy and ethics need to be implemented since the data user has different vested interests than the data providers.

Figure 6 shows the data analyst as a member of one of a pair of collaborating data provider organisations, and thus illustrates that the same organisation can be both a data provider and a data user at the same time.

Figure 7 shows an extension of this, where all collaborating organisations provide and use data.

How could this approach work to accelerate existing data sharing workflows?

The key driver for the Social Data Foundation is to find a way in which to accelerate responsible data access, collaboration and (re-)usage across the Council, Hospital and University, as well as other organisations, to enable positive health and social care transformation. We believe that the establishment of the Social Data Foundation as a TTPI would be able to accelerate existing data sharing workflows as follows:

(1) Better data discoverability

A key purpose for the Social Data Foundation is to *"match supply and demand between data suppliers and users"* as well as overcome information asymmetries between stakeholders (Richter & Slowinski, 2019), including data providers, data users, patients, service-users and citizens. A core function of the Social Data Foundation therefore is to enable greater data discoverability through a metadata catalogue. This would contribute to the acceleration of existing data sharing workflows for all key stakeholders as follows:

- Data users would have a better understanding about the types of data available and their utility through access to quality provenance metadata supplied by data providers – this would therefore reduce the need for speculative data access requests.
- It has the potential to attract further data providers to share data.
- It may help to reveal gaps, e.g. what data are not being collected, or what data are not made available for re-use.
- The use of standardised metadata would make it possible to potential join up with other catalogues, (e.g. those that are regional or national).
- It would offer greater visibility for citizens, including data subjects.

(2) Local solutions with national leadership

The Social Data Foundation would offer a localised hub for data-intensive research and innovation able to accelerate multi-party data sharing by establishing a strong presence and network of key stakeholders. Consequently, stakeholders can work together to discover solutions to health and social care transformation, promote greater collaboration, address key local priorities and rapidly respond to new and emerging health data-related challenges, whilst offering national exemplars of health system solutions.

(3) Empowering citizens

The Social Data Foundation would empower citizens by widening the range of stakeholders involved in key decision-making processes (data governance, design, evaluation, etc.) to include providers, civil society, and communities. All stakeholders would be better informed about needs and expectations increasing likelihood of data sharing, participation, and successful adoption of proposed changes.

(4) Greater assurances for stakeholders

The Social Data Foundation would provide greater assurances to stakeholders that best practice for data governance is followed, which will build trust and confidence in its operations. In particular, it would incentivise data providers to participate as data (re-)usage is monitored, which would help to accelerate data sharing as more health and social care data are likely to be made discoverable and accessible by existing and new data providers.

(5) Faster ethical oversight and information governance

It is imperative that all applicable ethics approvals are granted for each planned and that these processes are executed efficiently. As a TTPI, the Social Data Foundation would offer semi-automated business processes to rapidly establish approval requests, risk assessment (e.g. de-identification standards) and platform data-flows necessary for institutional and national approval requests (e.g. NHS HRA, NHS REC, Confidentiality Advisory Group (CAG)[29]).

How can this be achieved in practice?

The first crucial steps towards a workable Data Foundation are the design of a data governance structure combined with a suitable platform deployment scenario. We therefore need to identify:

- **THE STAKEHOLDERS** within the Social Data Foundation, including their individual roles and interests.
- **THE CORE DATA-RELATED FUNCTIONS** for the Social Data Foundation to operate effectively and appropriately – and ultimately deliver its mission for positive health and social care transformation.
- **THE DATA AND MANAGEMENT SERVICES OPERATED BY THE SOCIAL DATA FOUNDATION** through a trustworthy data sharing platform.

(1) Identify the stakeholders

The following nine core roles are required for the effective operation of the Social Data Foundation:

	SOCIAL DATA FOUNDATION: CORE ROLES
1	**Advisory Committee** A group of individuals external to the Social Data Foundation – with a wide range of expertise related to health and social care transformation (e.g. health and social care services, cyber-security, data governance, health data science, ethics, law) – that provides advice to the Social Data Foundation Board on matters related to data sharing (as necessary).[30]
2	**Citizen Representative** An expert in data governance, who is a mandatory member of the Social Data Foundation Board (see below) and oversees the administration of citizen participation and engagement activities to ensure that the Social Data Foundation maintains stakeholder approvals. In particular, the Citizen Representative shall (i) contribute to the definition of use cases as well as (ii) create and manage a framework for citizen participation and engagement activities.
3	**Data Provider** An entity who makes available specified health and social care data for (re-)usage by one or more data users as part of the Social Data Foundation. A representative of a data provider could act as a member of the Social Data Foundation Board.

[29] Pursuant to the common law duty of confidentiality, typically, the disclosure of confidential patient information should only transpire where the person to whom the information relates gives their consent (NHS HRA, 2018). However, for many research and innovation projects, obtaining consent would be impractical (e.g. for secondary use). S.251 of the National Health Service Act 2006 provides a 'statutory gateway' (The Information Governance Review, 2013) whereby confidential patient information can be disclosed for medical research (direct and indirect care) where the use of anonymous information is not possible and obtaining consent would be unfeasible (NHS HRA, 2018). The role of the Confidentiality Advisory Group therefore is to provide independent advice on whether applications for (re)use of confidential patient information for research or non-research purposes should be approved (NHS HRA, 2018; NHS HRA, 2020; The Information Governance Review, 2013).

[30] For example, the role of advisory committees have been outlined as part of potential basic governance structures for data trusts (Reed, C., BPE Solicitors & Pinsent Masons, 2019).

4 Data User

An entity who discovers, uses and/or re-uses shared data made accessible via the Social Data Foundation. It is important to note that the same organisation can be both a data user and a data provider in the same project. Note that a data analyst is a person who works for a data user and carries out data linking, query and analysis. A representative of a data user could act as a member of the Social Data Foundation Board.

5 Data Protection Officer

A standard role (whose appointment in some instances is mandatory under the GDPR) for organisations who process personal data to oversee the processing to ensure that it is compliant with GDPR obligations and respects data subjects' rights. For the Social Data Foundation, the DPO is responsible for overseeing the processing of any personal data within the Social Data Foundation and advising on compliance with the GDPR and the implementation of controls to address the risk of re-identification when data providers' data is linked in response to data users' queries. The DPO works closely with the Independent Guardian who is responsible for overseeing the processing of all types of data.

6 External Auditor

A body independent to the Social Data Foundation who is responsible for auditing or certifying its performance, conformance to standards and/or compliance to regulations.

7 Independent Guardian

A team of experts in data governance, who are independent from the Social Data Foundation Board and oversee the administration of the Social Data Foundation to ensure it achieves its purposes in accordance with its rulebook i.e. that all data related activities realise the highest standards of excellence for data governance. In particular, the Independent Guardian shall (i) help set up a risk-based framework for data sharing, (ii) assess the use cases in accordance with this risk-based framework and (iii) audit and monitor day-to-day all data-related activities, including data access, citizen participation and engagement.

8 Platform Facilitator

An executing officer, usually supported by a team, who oversees the technical day-to-day operation of the Trustworthy Data Sharing Platform,[31] including: the provision of infrastructure and functional services for data providers and data users, an access point for data users, and support services for other roles where required; the implementation of governance policies; and the hosting of a data repository if required.

9 The Social Data Foundation Board

The Social Data Foundation Board is an inclusive decision-making body whose appointed members represent the interests of stakeholders – data providers, data users and citizens – and therefore must include a mandatory Citizen Representative. The principal responsibility of its members is to administer the Social Data Foundation's assets and carry out its objects, including the determination of objectives, scope & guiding principles as well as governance policies & regulations through the comprehensive rulebook. All members shall carry out their duties both lawfully and ethically.

It is important to note that the sustainability and performance of the Social Data Foundation will be contingent on collaborative decision-making processes.[32]

The roles interact as shown in Figure 8.

(2) Ascertain the core data-related functions

The following table provides an overview of the core data-related functions to facilitate collaborative data sharing (i.e. operations and services) that could be provided by the Social

[31] Note that for the purposes of this white paper, we define 'Trustworthy Data Sharing Platform' as follows: The trust-enhancing technical and organisational infrastructure provided by the Data Foundation, which has potential to offer a range of data hosting and support services, and therefore enable responsible and trustworthy data discoverability, sharing, usage and re-usage.

[32] For background information on collaborative decision-making, see the "Decision Making Spectrum" (Goldminz, 2018). Also see Khatri & Brown (2010) who outline five key "interrelated decision domains" for data governance: "data principles", "data quality", "metadata", "data access" and "data lifecycle".

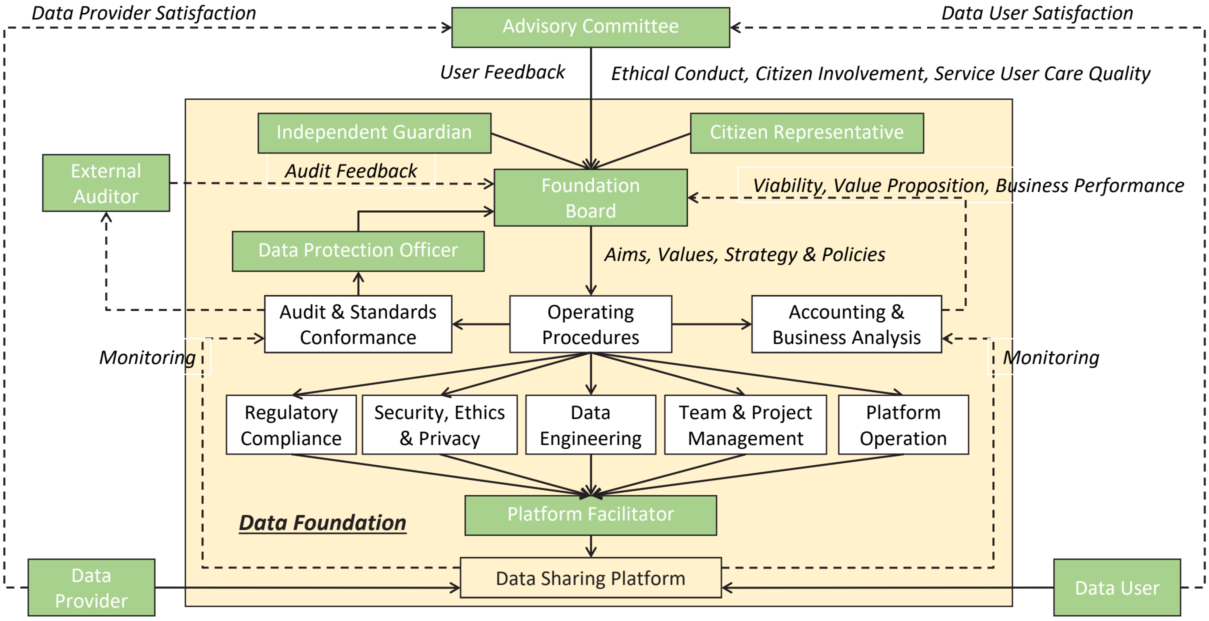


Figure 8. Social Data Foundation governance and management

Data Foundation's trustworthy data-sharing platform.

OVERVIEW OF CORE-DATA RELATED FUNTIONS TO FACILLITATE COLLABORATIVE DATA SHARING ORGANISED BY GROUP

GROUP	FUNCTION	DESCRIPTION
A. User Functions	*Register*	Data user registers at the Data Foundation.
	Find Data	Data user searches for data, e.g. using keywords. Relevant datasets returned.
B. Data Hosting	*Host Data*	Operate a data repository and store data provided by one or more data providers securely within it.
	Curate Data	Manage and maintain data over its lifetime stored within the data repository.
	Serve Metadata	Make metadata available to authorised users so that data may be discovered.
	Register Dataset Metadata	Data provider sends metadata of a hosted dataset to the Data Foundation so that the dataset may be discovered.
	Accept Data Access Requests	Operate an endpoint whereby users can request access.
	Enforce Access Policy	Determine whether access requests are granted or denied based on access policy for the requesting user and the requested dataset.
	Serve Data	Make data available to authorised users.
	Usage Audit	Securely record and store audit records of all data access by all users.
C. Data Preparation	*De-Identify Data*	Pre-process data to remove personal data.
	Annotate Data for Discovery	Enable data to be discovered in searches via annotation with e.g. keywords.
	Annotate Data for Linking	Enable data to be linked in analytics via e.g. semantic annotations referencing stated ontologies.
	Set Access Policy	Determine and state the access policy for a dataset to be applied when the data is requested by a data user.
D. Analysis	*Data Linking Query Processing*	Process queries that link multiple datasets together.
	Data Analytics	Perform analytical tasks on single or linked datasets.
	Monitor Linked Data for Personal Data	Monitor linked datasets for re-identification (personal data arising from the data linking), raise alarms and block

		release of data when detected.
	Query-Based Anonymisation/De-Identification	When linking data in response to a data user query, monitor the query result and take steps to ensure that the re-identification risk has been mitigated to an acceptable level considering both controls applied on the data and its environment.
E. Governance	*Determine Mission*	Determine values, objectives, purpose, and target beneficiaries.
	Determine Strategy	Determine how to achieve the mission, including high-level steering of the Data Foundation
	Determine Policies	Determine policies and regulations that set the rules of the Data Foundation.
	Determine Agreements	Determine the agreements needed between the stakeholders in the Data Foundation federation, their form and terms.
	Sign Agreements	Sign agreements with other stakeholders, e.g. data providers, data users.
	Identify Relevant Standards and Regulations	Identify relevant standards for the operation of the Data Foundation federation and incorporate them into policy.
	Regulatory Compliance and Standards Conformance	Ensure compliance to relevant regulation (e.g. GDPR), and certified conformance to relevant standards (e.g. UK Cyber Essentials Plus).
F. Support	*User Account Management*	Maintain a database of data user accounts.
	Host User Portal	Host a portal within a website where data users can find and access datasets.
	Search Support	Provide a search function that enables data users to search and find relevant datasets, e.g. using keywords.

(3) Determine the distribution of function

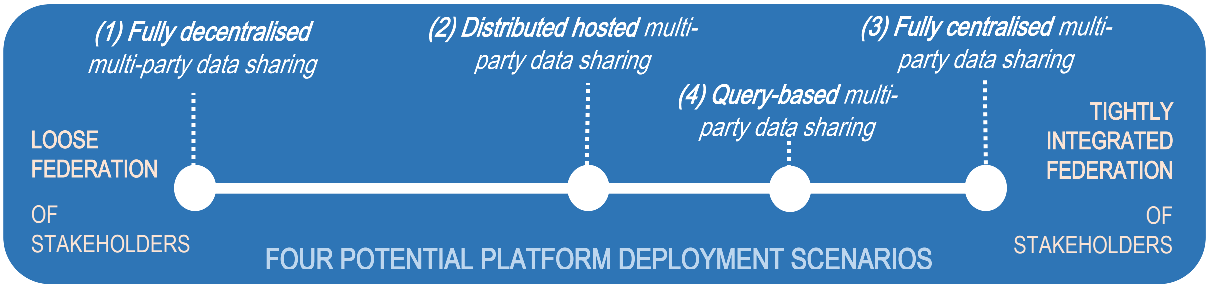

(1) Fully decentralised multi-party data sharing
(2) Distributed hosted multi-party data sharing
(3) Fully centralised multi-party data sharing
(4) Query-based multi-party data sharing
LOOSE FEDERATION
OF STAKEHOLDERS
TIGHTLY INTEGRATED FEDERATION
OF STAKEHOLDERS
FOUR POTENTIAL PLATFORM DEPLOYMENT SCENARIOS

Figure 9. Spectrum of federation options: Four platform deployment scenarios for the Social Data Foundation

The Social Data Foundation can be viewed as a federation of stakeholders each with varying degrees of authority and/or influence over decision-making processes.[33] Core data-related functions can be arranged within the federation in numerous ways – e.g. in one distribution of function, each data provider has full control over their data access policies, whereas in another an inclusive decision-making body can be delegated to make decisions over such policies.

The precise distribution of function therefore is directly linked to governance and risk, as it affects the dispersal of control, agency, and trust between stakeholders.

Four Platform Deployment Scenarios

The data governance structure is contingent on determining the most appropriate distribution of function for the Social Data Foundation. We have devised four platform deployment scenarios that span the spectrum of federation options from loose to tightly integrated federations of stakeholders, depicted in Figure 9.[34]

Platform Deployment Scenario 1: fully decentralised multi-party data sharing

Deployment Scenario 1 represents a **loose federation** of stakeholders where data providers are individually responsible for all data-related functions for their data sharing and use, including search and data hosting. To access data, data users must make individual requests to each data provider as applicable and establish bi-lateral agreements.

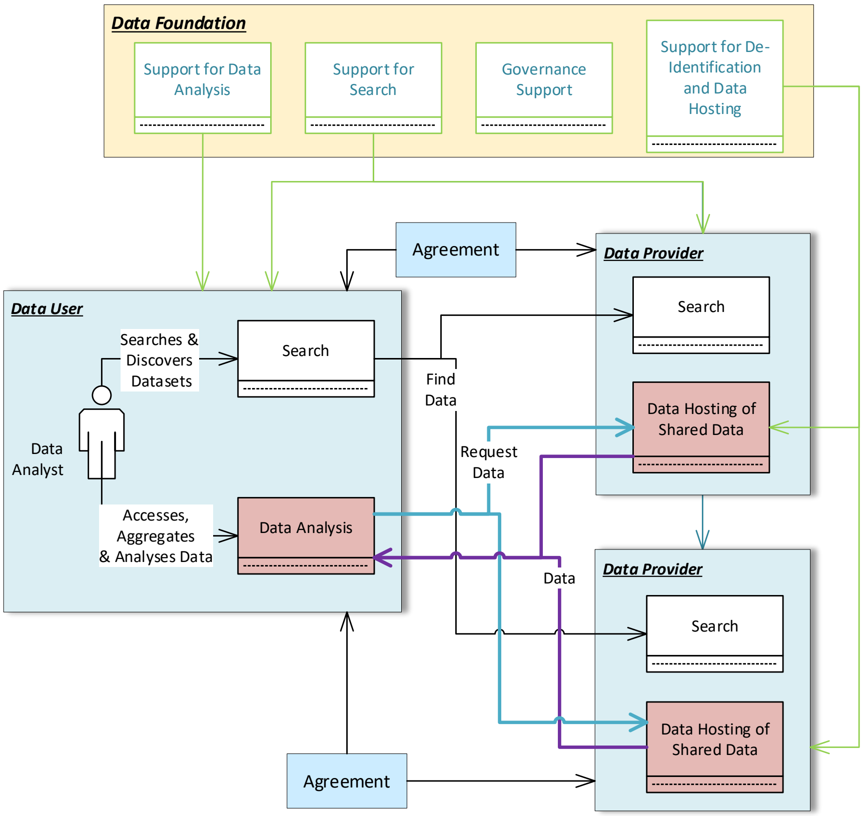

Data Foundation
Support for Data Analysis
Support for Search
Governance Support
Support for De-Identification and Data Hosting
Agreement
Data Provider
Search
Data Hosting of Shared Data
Data User
Searches & Discovers Datasets
Search
Data Analyst
Find Data
Request Data
Accesses, Aggregates & Analyses Data
Data Analysis
Data
Data Provider
Search
Data Hosting of Shared Data
Agreement

Figure 10. Platform Deployment Scenario 1: fully decentralised multi-party data sharing

[33] E.g. federated service management methodology (FitSM).

[34] Note that we have devised these four deployment scenarios based on the FitSM lightweight service management standard (FitSM; FitSM Expert), which explicitly addresses different types of federation across a spectrum of looser to more tightly integrated federations.

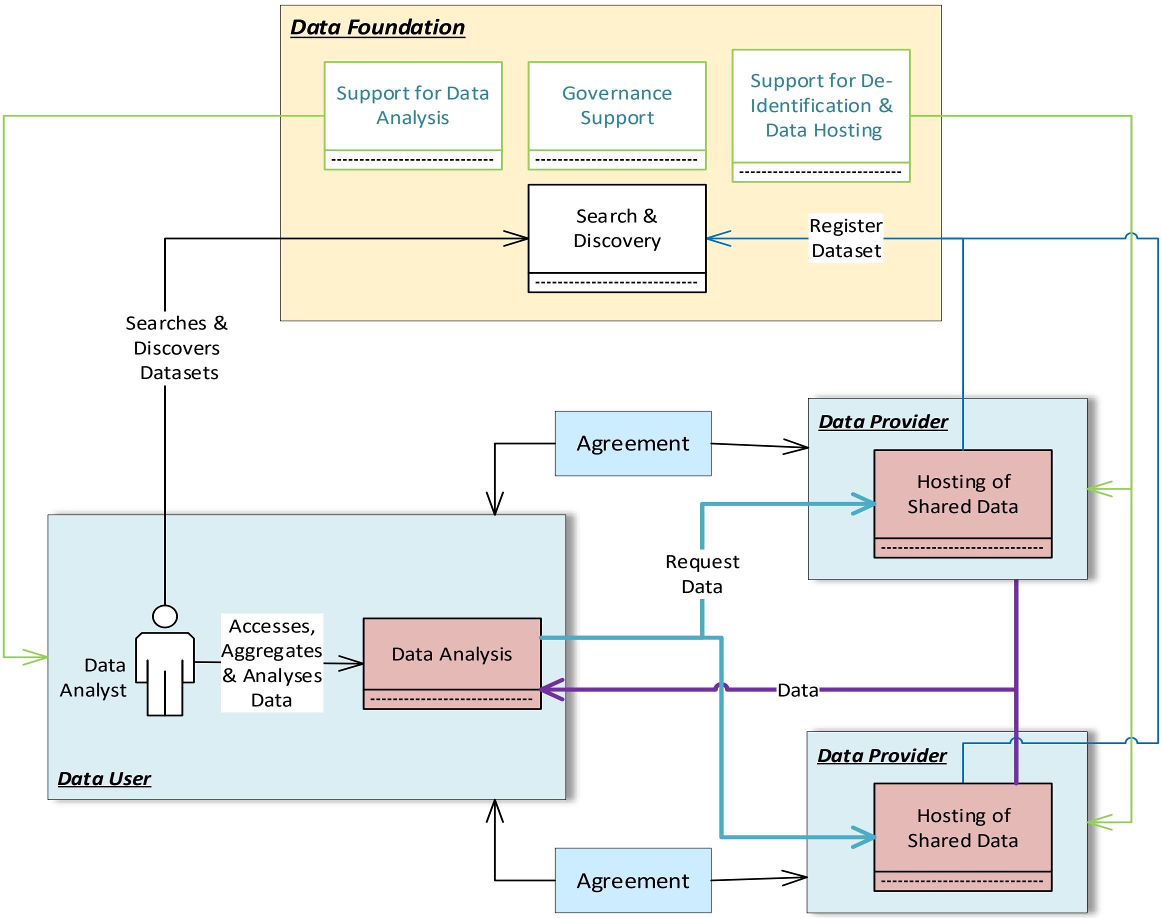


Figure 11. Platform Deployment Scenario 2: distributed hosted multi-party data sharing

In a fully decentralised scenario, **the role of the TTPI is limited to a support service provider.** For instance, possible technical support, mentoring and consultancy for data provider(s) and data user(s) where applicable or data governance support, e.g. as a standards/guidance body, certification body.

Platform Deployment Scenario 2: distributed hosted multi-party data sharing

Deployment Scenario 2 represents a **moderately integrated federation**[35] of stakeholders where data providers are individually responsible for data hosting and data preparation functions, and are required to make their individual datasets safe and useful.

To access data, data users must make individual requests to each data provider as applicable. However, in this scenario, **the TTPI provides a centralised search and discovery service for all shared data**, in addition to other technical and governance support services.

Platform Deployment Scenario 3: centralised multi-party data sharing

Deployment Scenario 3 represents a **tightly integrated federation**[36] of core stakeholders where shared data are sub-licensed to the trusted third-party intermediary for re-use by data users. Shared data and metadata are anonymised or de-identified by data providers

[35] Note that deployment scenario 2 is most closely aligned with the matchmaker federation type (FitSM Expert).

[36] Note that this centralised deployment scenario is most closely aligned with the centralised federation type (FitSM Expert).

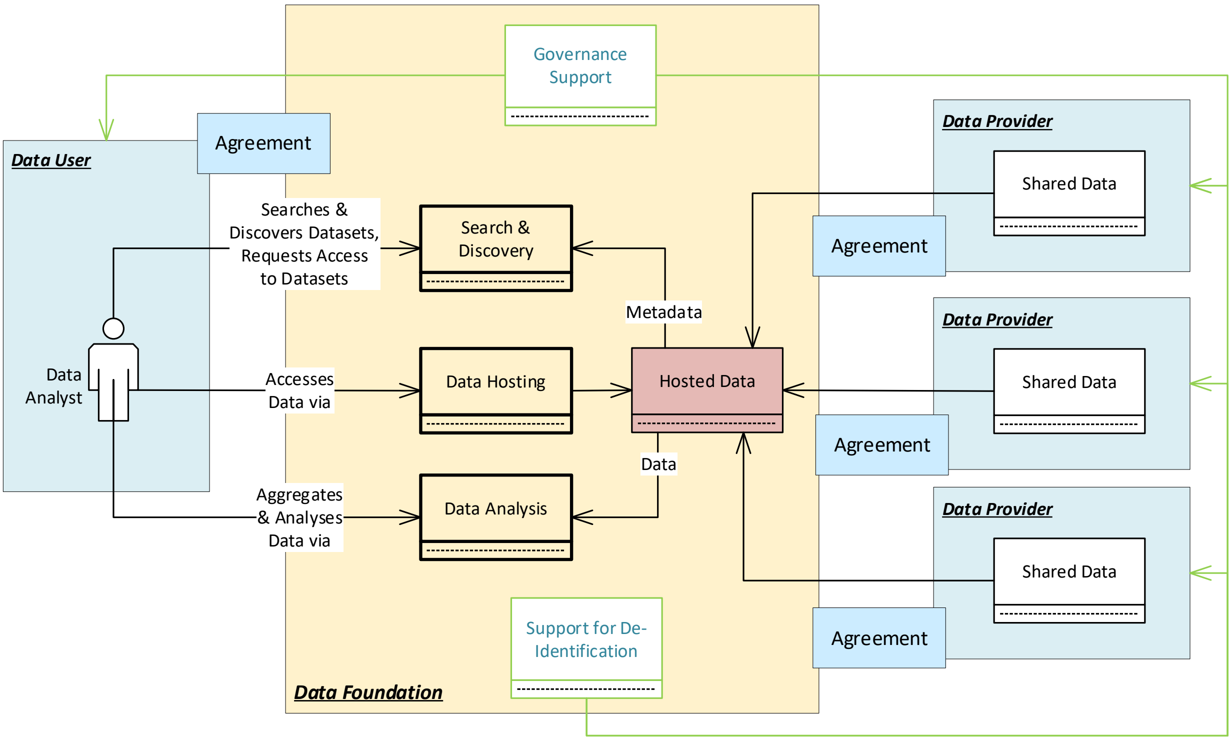


Figure 12. Platform Deployment Scenario 3: centralised multi-party data sharing

(supported by the TTPI) and transmitted to the trustworthy data sharing platform.

To access data, data users can both discover relevant data and make requests to access one or more shared datasets. In this scenario, **the TTPI is a collective decision-maker** with a high level of oversight over day-to-day data-related activities**.** In addition to centralised search and discovery services as well as other technical and governance support services, it hosts shared data and can sub-license shared data, i.e. can make independent decisions to reject or accept potential projects.

In this scenario, **significant trust is placed in the TTPI by data providers**, as the TTPI holds and manages access to shared data. There is therefore a need to provide assurances to data providers, in particular that: (i) data provided is being managed correctly by the TTPI; (ii) the trustworthy data sharing platform is secure and resilient to cyber threats; and (iii) the TTPI enforces data access in line with the specified requirements of the data provider. Data sharing agreements are vital (e.g. to set out liabilities and responsibilities) as well as independent governance and audit to ensure that the TTPI is adhering to these agreements.

Given the data provider is unlikely to know beforehand who wishes to access its data held within the Social Data Foundation, there is a need for a dynamic aspect to data sharing agreements and data access policies. For instance, flexible policies for data access (e.g. pre-defining roles or groups who are eligible to access shared data as well as limits) and protocols (e.g. the TTPI must ask the data provider for permission every time a request is made for their data) need to be investigated and selected based on the particular set of circumstances and the needs of the stakeholders.

A crucial benefit of this deployment scenario is that once data is cleaned, prepared for linkage and stored within the TTPI, it can be retained within the TTPI **ready for access by**

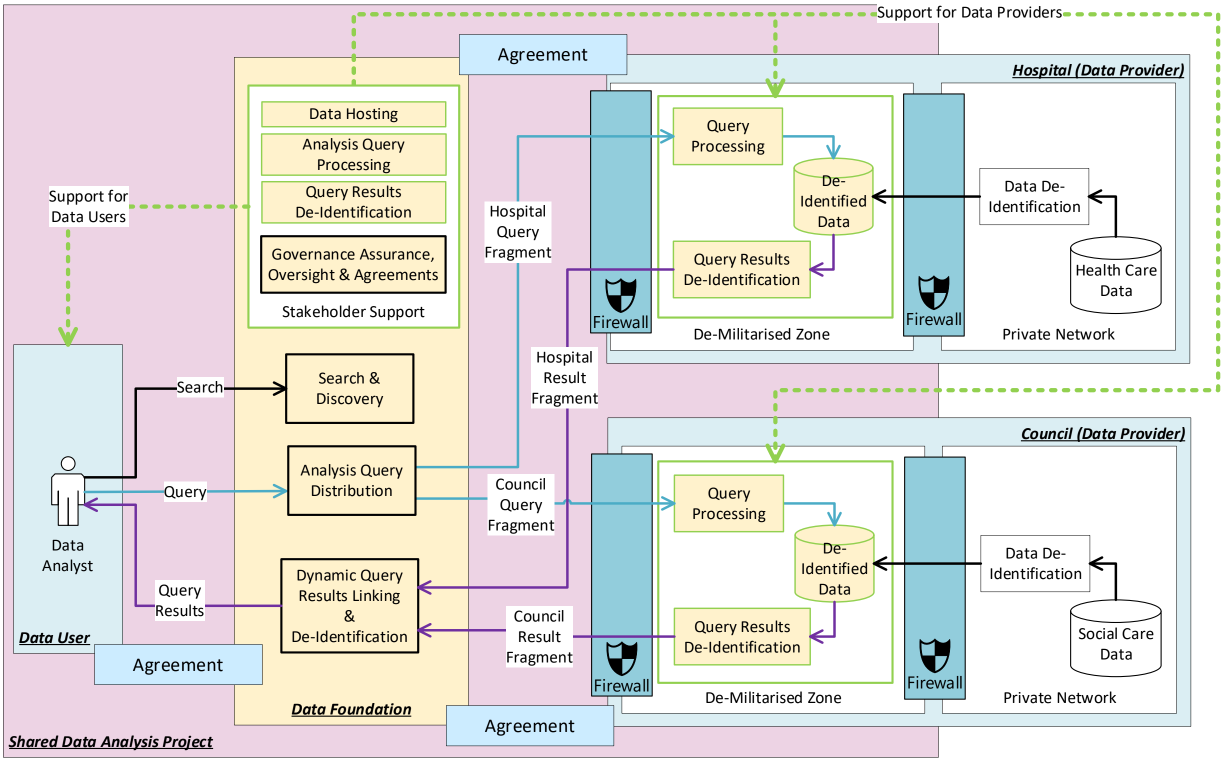


Figure 13. Platform Deployment Scenario 4: query-based multi-party data sharing initiative

authorised users without the need to rely on technical engagement from data providers or other parties. Furthermore, there is potential for the TTPI to offer a safe, secure, and ethical service for dynamic linking for data providers and data users.

Platform Deployment Scenario 4: query-based multi-party data sharing

Deployment Scenario 4 represents another **tightly integrated federation** of core stakeholders where shared data are accessible to data users through a linked data query processing service. This scenario differs from the previous in that only the results of pre-agreed data analysis queries are shared beyond the boundaries of data providers. In this scenario, the TTPI facilitates projects, principally, through:

i. Orchestration of data search, discovery, querying of linked data from multiple providers and security context management for data users.

ii. Governance, agreements, shared data management and assurance services for data providers and users.

iii. Tools and services to implement secure sandboxes that enable sharing operations (e.g. de-identification and query processing) at data providers.

Figure 13 provides a high-level overview of a query-based multi-party data sharing initiative – note that the Hospital and Council are utilised as two examples of possible data providers. The six key operating principles of the TTPI are as follows:

- **PRINCIPLE 1 – The Social Data Foundation acts as a trusted third party intermediary (TTPI) to facilitate shared data analysis projects** via governance, brokerage of agreements between data providers and data users, shared data management and assurance services, a front-end portal, and tooling to enable sharing operations that are executed at data providers (e.g. for de-identification).

- **PRINCIPLE 2 – The Social Data Foundation provides a dynamic linking service** for

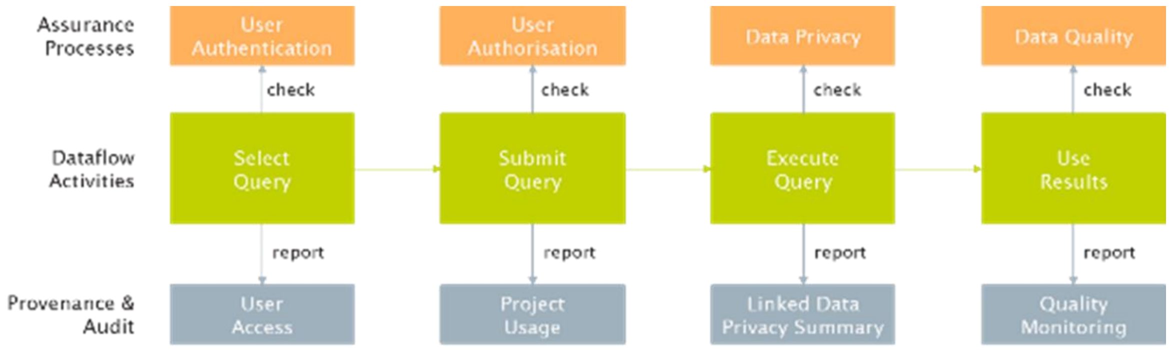


Figure 14: Workflow for an authorised query submitted from a data user perspective to the TTPI

authorised data users where two or more sources of health and social care data are brought together on demand according to the specific parameters of an authorised data user's query where the risk of re-identification is both evaluated before and after data linkage, and mitigated through assurance processes facilitated by the Data Foundation.

- **PRINCIPLE 3 – The extent of "data sharing" is limited to the results of pre-approved queries agreed by all parties in a project – not whole datasets.** The Data Foundation facilitates a process to approve queries based on a risk assessment and provides a gateway for data analysis queries from authorised data users.

- **PRINCIPLE 4 – The Social Data Foundation carries out a risk assessment for each shared data analysis project before any data is shared** by data providers and assigns a list of pre-approved queries to authorised data users.

- **PRINCIPLE 5 – Data providers only share de-identified data as part of the Social Data Foundation.** The possible risk of re-identification – related to a specific pre-approved data analysis query – is addressed at the point of delivery by each data provider before their data is linked with other data providers' data, as well as at the point of linking at the Data Foundation and mitigated through assurance processes facilitated by the Data Foundation.

- **PRINCIPLE 6 – Agreements govern relationships between all stakeholders** for each shared data analysis project, including the assignment of pre-approved queries to one or more authorised data users as part of a specific project.

How would this scenario work from a data user perspective?

Prospective data users would register with the TTPI and request a health and social care project(s). The TTPI would review the request and if approved, the project would be assigned a set of pre-approved queries. The queries would form part of a data sharing agreement and would allow data users who are members of the project to link shared data across one or more data providers. Data users then would be able to submit and receive responses to their queries through the data sharing platform provided by the TTPI. In this scenario, the extent of "data sharing" therefore is limited to the results of pre-approved queries – not underlying datasets.

For illustration, Figure 14 provides an overview of a workflow for an authorised query submitted by a registered data user.

How would this scenario work from a data provider perspective?

Data providers would sub-license their de-identified data to the TTPI to permit the TTPI to provide responses to queries concerning data linked from multiple providers for registered data users. Data providers therefore only share de-identified data with the TTPI. Given this scenario is focused on dynamic data linkage, there would be a robust three-layer approach for de-identification in place:

QUERY-BASED SCENARIO: THREE LAYER APPROACH FOR DE-IDENTIFICATION

- **Layer 1. Internal de-identification.** Data providers de-identify their data sources within their private network (i.e. static stage of redaction). The results of this de-identification process are then transmitted to a secure area that is segregated from their private network (known as their de-militarised zone – DMZ).
- **Layer 2. Query-based anonymisation/de-identification.** A data user requests a pre-approved query linking data from multiple providers via the TTPI. The TTPI forwards query fragments onto each relevant data provider. The de-identified source data within the DMZ(s) of the data provider (s) are further anonymised or de-identified in accordance with the context and purpose of the specific query and a result set is returned to the TTPI.
- **Layer 3. De-identification of linked data.** The TTPI links each result fragment from providers together and the combined linked result is checked for the risk of re-identification. Only if an acceptable, low level of risk is found will the results be presented to the data user – e.g. the risk of re-identification is considered to be no more than remote. If an unacceptable level of risk is found, the data is not released to the data user pending further checking and additional measures to de-identify it.

Summary: four Deployment Scenarios

Based on our analysis, we do not consider platform deployment scenarios 1 and 2 as viable options for a Social Data Foundation. The fully decentralised scenario presented is incompatible with the rationale for the Data Foundation to independently steward data and facilitate greater citizen participation and engagement. Furthermore, aside from greater data discovery, and technical and governance support services, the distributed hosted scenario outlined does not add sufficient value to existing data sharing ecosystems and practices or accelerate data sharing for health and social care transformation.

A Social Data Foundation requires a distribution of function that affords a high level of collaborative decision-making, oversight of day-to-day data sharing practices and gives rise to a data governance structure supported by platform deployment scenarios 3 and 4.

Based on our analysis, we consider the centralised and query-based scenarios to be possible options for deployment. As with the centralised scenario, the distribution of function for the query-based scenario is well suited to the requirements of the data foundation framework – in particular, there is opportunity for: (i) independent data stewardship, (ii) data sharing advocacy, (iii) accelerated data sharing, usage and re-usage, and (iv) a high degree of collaborative decision-making and oversight over data-related activities.

Both the centralised and query-based scenarios could enable dynamic linking. A crucial benefit of dynamic linking is that the privacy-utility trade-off can be assessed at a very granular level – i.e. per project – to establish the optimal level of utility whilst preserving privacy through targeted technical and organisational measures. With greater flexibility to regulate the privacy-utility trade-off, this is likely to increase and/or accelerate data linkage in contrast to other data sharing initiatives with more generalised privacy-preserving polices and controls.

A critical advantage of the query-based approach is that data providers retain greater control over their datasets, adding the value of dynamic linking to existing data processes within an assured environment, whilst minimising replication, retention periods and associated costs.

Whereas a potential disadvantage of the centralised approach is that some data providers may be reticent to provide their de-identified data to a central repository where they have less control, or are unable to do this, e.g. where certain data can only be accessed on site.

SOME KEY OBSERVATIONS ON TIME AND EFFORT REQUIRED TO MANAGE THIRD PARTY DATA FOR SAFE LINKAGE[37]

We recognise that acquiring new data providers, as well as further datasets from existing data providers, will involve significant time and effort from a Social Data Foundation – in particular, preparation for data linkage as well as handling legal arrangements and licensing agreements with multiple data providers and data users.

(1) Ensuring all shared data can be integrated for linkage

Each new dataset provided as part of the Data Foundation will need to be prepared for integration with existing linked datasets. Such data preparation tasks include but are not limited to: (i) data de-identification; (ii) data cleaning; (iii) data quality assurance; (iv) data consistency assurance (e.g. ensuring pseudonymised identifiers are consistent across datasets); and (v) data compatibility assurance (e.g. normalising data fields across heterogeneous data sets generated by different software).

(2) Handling legal arrangements, licensing agreements, and rights management & clearance

Legal arrangements and licensing agreements also require time and effort to draft, negotiate, verify and (where applicable – e.g. a contract) sign. Due diligence must ensure intellectual property rights management and clearance, and effective and appropriate procedures in place to manage (sensitive) personal data. **For robustness and efficiency, a Social Data Foundation will require standardised protocols and procedures for data preparation, contracts management and due diligence – as well as standards for legal arrangements and licensing agreements (as applicable). Exploration will be necessary to determine how processes could be semi-automated in the medium to long-term using advances in smart contracting technologies.**

[37] We greatly acknowledge our valuable discussions with Privitar on which these observations are based. Please note that all views and opinions expressed are those of the authors.

5. Conclusion

Our goal is to foster a trustworthy data sharing alliance between the Southampton City Council, the University Hospital Southampton NHS Foundation Trust and the University of Southampton, as well as flexible membership to other organisations through the establishment of a Social Data Foundation.

We believe that a Social Data Foundation, as a TTPI, would be able to accelerate multi-party data sharing for health and social care transformation as follows:

THE PRINCIPAL WAYS A SOCIAL DATA FOUNDATION COULD ACCELERATE DATA SHARING:

- ✓ **BETTER DATA DISCOVERABILITY.** Through a comprehensive metadata catalogue, data users and citizens would have a better understanding about the data available and utility through quality provenance metadata supplied by data providers.
- ✓ **LOCAL SOLUTIONS WITH NATIONAL LEADERSHIP.** As a localised hub for data-intensive research and innovation and positive health and social care transformation, a Social Data Foundation would be able to promote greater collaboration, address key local priorities and rapidly respond to new and emerging health data-related challenges, whilst offering national exemplars of health system solutions.
- ✓ **EMPOWERING CITIZENS TO CO-CREATE AND PARTICIPATE IN SYSTEMS TRANSFORMATION.** By widening the range of stakeholders involved in key decision-making processes (data governance, design, evaluation, etc.) to include providers, civil society and communities all stakeholders will be better informed about needs and expectations increasing likelihood of data sharing, participation and successful adoption of proposed changes.
- ✓ **GREATER ASSURANCES THAT BEST PRACTICE DATA GOVERNANCE IS FOLLOWED.** Building trust and confidence between stakeholders in Data Foundation operations is necessary to ensure safe and useful data sharing.
- ✓ **FASTER ETHICAL OVERSIGHT AND INFORMATION GOVERNANCE.** As a TTPI, a Social Data Foundation would offer semi-automated business processes to rapidly establish approval requests, risk assessment (e.g. de-identification standards) and platform data-flows necessary for institutional and national approval requests (e.g. NHS HRA, NHS REC or CAG).

Recommended Platform Deployment Scenario

Based on our analysis, we consider the centralised and query-based scenarios to be the most well suited options for deployment for the following key reasons:

- ✓ **HIGH PLATFORM COLLABORATION** through collective decision-making body with influence over data discovery, access, and controls.
- ✓ **HIGH PLATFORM UTILITY** through involvement in core data preparation and data query functions.
- ✓ **HIGH DATA FINDABILITY** through data discovery for all data shared by data providers.
- ✓ **HIGH DATA ACCESSIBILITY** through a single data query made directly for one or more shared datasets held by several data providers.
- ✓ **HIGH DATA ASSURANCE** through data functions executed either in collaboration or independently from data providers with accountable oversight, maximising data stewardship and data sharing advocacy.
- ✓ **DYNAMIC LINKING** allows for more a granular approach to the utility-privacy trade-off to establish the optimal level of utility for each shared data analysis project whilst preserving privacy through targeted technical and organisational measures.

While we recognise that a centralised approach may be a more realistic option for the immediate operation of a Social Data Foundation, we recommend that work to advance towards a query-based model should begin from the outset.

The query-based approach is preferable as there is minimised data replication, retention and associated costs, as data is not stored centrally beyond the needs of specific projects at the point of use, with potential for "caching" and reconstruction.

Moving forward, our ambition is that the establishment of a Social Data Foundation for positive health and social care transformation acts as a springboard to accelerate trustworthy and collaborative data sharing within and across other domains. Ultimately, we hope this leads to a wider network of social data foundations connected by a shared and advancing knowledge base of best practice for safe data linkage and governance.

Acknowledgments

The Social Data Foundation Project is partly funded and supported by the University of Southampton’s Web Science Institute (WSI) and Southampton Connect.

We extend our special thanks to Rachel Bailey (UHS), Tom Barnett (WSI, UoS), Prof Sally Brailsford (CORMSIS, UoS), Guy Cohen & Marcus Grazette (PRIVITAR), Christine Currie (CORMISIS, UoS), Jo Dixon (RIS, UoS), Alison Knight (RIS, UoS), Alistair Sackley (WSI, UoS), Prof Mike Surridge (IT Innovation, UoS), Neil Tape (UHS), and Gary Todd (FAMIIO).

We also greatly acknowledge the valuable discussions and feedback from PRIVITAR on an earlier draft of this white paper.

Please note that all views and opinions expressed in this white paper are those of the authors, and do not necessarily represent those named above.

References

Aitken, M. et al. (2019). Consensus Statement on Public Involvement and Engagement with Data-Intensive Health Research. International Journal of Population Data Science, 4(1). Retrieved from: https://ijpds.org/article/view/586.

Alexin, Z. (2014). Does fair anonymization exist? International Review of Law, Computers & Technology, 28(1), 21-44, Retrieved from: https://www.tandfonline.com/doi/full/10.1080/13600869.2013.869909.

Arnstein, S. R. (1969). A Ladder of Citizen Participation. Journal of the American Institute of Planners, 35(4), 216-224. Retrieved from: https://doi.org/10.1080/01944366908977225.

Article 29 Data Protection Working Party. (2014, April 10). Opinion 05/2014 on Anonymisation Techniques. (0829/14/EN; WP216), Adopted on 10 April 2014, Retrieved from: http://www.pdpjournals.com/docs/88197.pdf.

CARE Principles for Indigenous Data Governance (2018, November 8). International Data Week and Research Data Alliance Plenary co-hosted event “Indigenous Data Sovereignty Principles for the Governance of Indigenous Data Workshop,” Gaborone, Botswana. Retrieved from: https://www.gida-global.org/care.

Carter, P., Laurie, G.T., & Dixon-Woods, M. (2015). The social licence for research: why *care.data* ran into trouble. Journal of Medical Ethics, 41(5), 404-409. Retrieved from: http://dx.doi.org/10.1136/medethics-2014-102374.

Cassell, A. et al. (2018). The epidemiology of multimorbidity in primary care: a retrospective cohort study. British Journal of General Practice, 68 (669), e245–51, Retrieved from https://doi.org/10.3399/bjgp18X695465.

CORMSIS (2020). Internal project correspondence.

CurvedThinking. (2019, July). Understanding public expectations of the use of health and care data. Developed in consultation with: Understanding Patient Data, Commissioned by OneLondon, Retrieved from: https://www.northlondonpartners.org.uk/ourplan/Areas-of-work/Digital/understanding-public-expectations-of-the-use-of-health-and-care-data.pdf.

Data Protection Act 2018 (UK). Retrieved from: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted.

Department of Health, the Caldicott Committee. (1997, December). Report on the Review of Patient-Identifiable Information. Retrieved from: https://webarchive.nationalarchives.gov.uk/20130124064947/http:/www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4068404.pdf.

Dodds, L. et al. (2020, April). Designing sustainable data institutions. Open Data Institute (ODI) report, Retrieved from: https://theodi.org/article/designing-sustainable-data-institutions-paper/.

Elliot, M. et al. (2016). The Anonymisation Decision-Making Framework. UKAN Publications, Retrieved from: https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf.

Elliot, M. et al. (2018). Functional anonymisation: Personal data and the data environment. Computer Law & Security Review, 34(2), 204-221, Retrieved from: https://doi.org/10.1016/j.clsr.2018.02.001.

FitSM Expert. Expert training in IT service management according to FitSM, v.1.4, Retrieved from: https://www.fitsm.eu/download/402/.

FitSM. Standards for lightweight IT Service Management. Developed by The FedSM Project, funded by the European Commission, Maintained by ITEMO, Retrieved from: https://www.fitsm.eu/fitsm-standard/.

General Data Protection Regulation (GDPR). Regulation (EU) 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/EC (General Data Protection Regulation). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679.

Goldminz, I. (2018, March 27). The Decision Making Spectrum. Medium, Retrieved from: https://medium.com/org-hacking/the-decision-making-spectrum-d8069d73a651.

Hardinges, J., & Tennison, J. (2020, February 10). What do we mean by data institutions? Open Data Institute (ODI) Blog, Retrieved from: https://theodi.org/article/what-do-we-mean-by-data-institutions/.

Health Data Research (HDR) UK. (2020a). Involvement and Engagement Guiding Principles. Retrieved from: https://www.hdruk.ac.uk/what-is-health-data-research/patient-and-public-involvement-and-engagement/patient-and-public-involvement-and-engagement-guiding-principles/.

Health Data Research (HDR) UK. (2020b). Opportunities to Get Involved. Retrieved from: https://www.hdruk.ac.uk/what-is-health-data-research/patient-and-public-involvement-and-engagement/opportunities-to-get-involved/.

Health Data Research Innovation Gateway. About. Retrieved from: https://www.healthdatagateway.org/pages/about.

Health Research Authority. Integrated Research Application System (IRAS). Retrieved from: https://www.myresearchproject.org.uk/.

Information Commissioner's Office. (2012). Anonymisation: managing data protection risk code of practice. Retrieved from: https://ico.org.uk/media/1061/anonymisation-code.pdf.

Involve. (2020a). NHS Citizen – Can citizens participate at the heart of NHS decision-making? Retrieved from: https://www.involve.org.uk/our-work/our-projects/practice/can-citizens-participate-heart-nhs-decision-making.

Involve. (2020b). Public Engagement Training. Retrieved from: https://www.involve.org.uk/our-work/public-engagement-training.

ISO. (2009a). ISO/IEC 15408-1:2009 Information technology -- Security techniques -- Evaluation criteria for IT security -- Part 1: Introduction and general model. Retrieved from: https://www.iso.org/.

ISO. (2009b). ISO/IEC 31000:2009 Risk management -- Principles and guidelines, 2009. Retrieved from: https://www.iso.org/.

ISO. (2011) ISO/IEC 27005:2011. Information technology -- Security techniques -- Information security risk management, International Organization for Standardization, 2011. Retrieved from: https://www.iso.org/standard/56742.html.

ISO. (2013a). ISO/IEC 27001:2013. Information technology – Security Techniques – Information security management systems – Requirements, International Organization for Standardization, 2013. Retrieved from: https://www.iso.org/.

ISO. (2013b). ISO/IEC 27002:2013 Information technology -- Security techniques -- Code of practice for information security management, 2013. Retrieved from: https://www.iso.org/standard/54533.html.

ISO. (2019). ISO, ISO/IEC 31010:2019 Risk management - Risk assessment techniques. Retrieved from: https://www.iso.org/standard/72140.html

Jones, K. (2020, April 1). SAIL Databank – Public Involvement & Engagement Policy. Retrieved from: https://saildatabank.com/wp-content/uploads/200416-Public-Involvement-Engagement-Policy.pdf.

Jones, K.H. et al. (2017). The other side of the coin. Harm due to the non-use of health-related data. International Journal of Medical Informatics, 97, 43-51. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1386505616302039.

Jones, K.H., & Ford, D.V. (2018). Population data science: advancing the safe use of population data for public benefit. Epidemiology and Health, 40, e2018061. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6367205/.

Khatri, V., & Brown, C.V. (2010). Designing Data Governance. Communications of the ACM, 53(1), 148-152, Retrieved from: https://cacm.acm.org/magazines/2010/1/55771-designing-data-governance/fulltext#T2.

Kyriazis, D., et al. (2017). CrowdHEALTH: Holistic Health Records and Big Data Analytics for Health Policy Making and Personalized Health. Informatics Empowers Healthcare Transformation, Stud Health Technol Inform, 238, 19-23, Retrieved from: https://pubmed.ncbi.nlm.nih.gov/28679877/.

Lin, D., et al. (2020). The TRUST Principles for Digital Repositories. Scientific Data, 7, 144, Retrieved from: https://www.nature.com/articles/s41597-020-0486-7.

Medicines & Healthcare Products Regulatory Agency. (2020, June 15). Consultation outcome – Response: What we will do differently. Retrieved from: https://www.gov.uk/government/consultations/how-should-we-engage-and-involve-patients-and-the-public-in-our-work/outcome/response-what-we-will-do-differently.

National Cyber Security Centre (UK). Cyber Essentials. Retrieved from: https://www.ncsc.gov.uk/cyberessentials/overview

National Health Service Act 2006. Retrieved from: https://www.legislation.gov.uk/ukpga/2006/41/contents.

National Institute for Health Research, Chief Scientist Office, Health and Care Research Wales and Public Health Agency. (2019). National Standards for Public Involvement. Supported by the Standards Development Partnership. Retrieved from: https://www.invo.org.uk/wp-content/uploads/2019/02/71110_A4_Public_Involvement_Standards_v4_WEB.pdf.

NHS Digital. Data Security and Protection Toolkit. Retrieved from: https://www.dsptoolkit.nhs.uk/.

NHS England. (2020). Clinical Networks and Clinical Senates. Retrieved from: https://www.england.nhs.uk/south-east/about-us/networks-senates/.

NHS Health Research Authority (NHS HRA). (2020, October 20). Approvals and amendments: Confidentiality Advisory Group (CAG). Retrieved from: https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/confidentiality-advisory-group/.

NHS Heath Research Authority (NHS HRA). (2018, May 9). Why is confidential patient information used? Retrieved from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/why-confidential-patient-information-used/.

NHS Heath Research Authority (NHS HRA). Confidentiality Advisory Group. Retrieved from: https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/.

NHS Scotland. Public Benefit and Privacy Panel for Health and Social Care. Retrieved from: https://www.informationgovernance.scot.nhs.uk/pbpphsc/.

Organisation for Economic Co-operation and Development (OECD). (2016, December 13). Recommendation of the Council on Health Governance – OECD/LEGAL/0433. Adopted on 13 December 2016, Retrieved from: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0433.

Organisation for Economic Co-operation and Development (OECD). (2019, November). Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies: Chapter 4. Risks and challenges of data access and sharing. OECD Publishing, Paris, Retrieved from: https://doi.org/10.1787/276aaca8-en.

Oswald, M. (2013) Something Bad Might Happen: Lawyers, Anonymization, and Risk. XRDS: Crossroads, The ACM Magazine for Students - The Complexities of Privacy and Anonymity, 20(1), 22-26, Retrieved from: https://dl.acm.org/doi/pdf/10.1145/2508970.

Public Health Research Data Forum. (2015, March). Enabling Data Linkage to Maximise the Value of Public Health Research Data. Final report to the Wellcome Trust, Retrieved from: https://wellcome.org/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf.

Public Participation Team. (2017, April). NHS England Patient and public participation policy. Version 2, Retrieved from: https://www.england.nhs.uk/wp-content/uploads/2017/04/ppp-policy.pdf.

Reed, C., BPE Solicitors & Pinsent Masons. (2019, April). Data trusts: legal and governance considerations. Retrieved from: https://theodi.org/wp-content/uploads/2019/04/General-legal-report-on-data-trust.pdf.

Richter H., & Slowinski, P.R. (2019). The Data Sharing Economy: On the Emergence of New Intermediaries. International Review of Intellectual Property and Competition Law, 50, 4-29, Retrieved from: https://doi.org/10.1007/s40319-018-00777-7.

Sane, J., & Edelstein, M. (2015, April). Overcoming Barriers to Data Sharing in Public Health: A Global Perspective. Research Paper, Centre on Global Health Security, Chatham House, The Royal Institute of International Affairs, Retrieved from: https://www.chathamhouse.org/sites/default/files/field/field_document/20150417OvercomingBarriersDataSharingPublicHealthSaneEdelstein.pdf.

Scott, K. et al. (2018, April). Data for Public Benefit: Balancing the risks and benefits of data sharing. Report Co-authored by Understanding Patient Data, Involve and Carnegie UK Trust, Retrieved from: https://www.involve.org.uk/sites/default/files/field/attachemnt/Data%20for%20Public%20Benefit%20Report_0.pdf.

Scottish Government. (2015, November 16). Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics. Retrieved from: https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/pages/5/.

Southampton City Five Year Health and Care Strategy. Retrieved from: https://hiowhealthandcare.org/application/files/2515/7527/8501/Southampton_Five_Year_Health__Care_Staretgy_HOSP2.pdf.

Southampton Connect. Southampton City Strategy 2015-2025. Retrieved from: https://www.southampton.gov.uk/images/southampton-city-strategy-15-25_tcm63-387730.pdf.

Stalla-Bourdillon, S., Carmichael, L., & Wintour, A. (2020, September). Fostering Trustworthy Data Sharing: Establishing Data Foundations in Practice. Data for Policy Conference 2020, Retrieved from: https://zenodo.org/record/3967690#.X2oCJ2hKg2w.

Stalla-Bourdillon, S., Wintour A., & Carmichael. L. (2019). Building Trust through Data Foundations: A Call for a Data Governance Model to Support Trustworthy Data Sharing. Web Science Institute (WSI) White Paper #2, Retrieved from: https://www.southampton.ac.uk/wsi/enterprise-and-impact/white-papers.page.

The Information Governance Review. (2013, March). Information: To share or not to share? Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf.

The National Archives. (2020). NHS Citizen archived webpages. Retrieved from: https://webarchive.nationalarchives.gov.uk/*/https:/www.nhscitizen.org.uk/.

The UK Caldicott Guardian Council. A Manual for Caldicott Guardians: The Caldicott Principles. Retrieved from: https://www.ukcgc.uk/manual/principles.

UK Data Service. Regulating access to data: 5 Safes. Retrieved from: https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes.

UK Health Data Research Alliance. (2020, April 30). Trusted Research Environments (TRE): A strategy to build public trust and meet changing health data science needs. Draft Green Paper v1.0 for consultation, Retrieved from: https://ukhealthdata.org/wp-content/uploads/2020/04/200430-TRE-Green-Paper-v1.pdf.

UK Standards for Public Involvement: Better public involvement for better health and social care research. National Institute for Health Research (NIHR), Chief Scientist Office, Ymchwil lechyd a Gofal Cymru/Health and Care Research Wales, Public Health Agency, Supported by the UK Public Involvement Standards Partnership, Retrieved from: https://sites.google.com/nihr.ac.uk/pi-standards/standards.

University of Southampton. Ethics and Research Governance Online (ERGO) 2. Retrieved from: https://www.ergo2.soton.ac.uk.

van Panhuis, W.G. et al. (2014). A systematic review of barriers to data sharing in public health. BMC Public Health, 14 (1144), Retrieved from: https://doi.org/10.1186/1471-2458-14-1144.

Appendices

Appendix A: Some examples of existing data sharing initiatives[38]

- Data Access Request Service (DARS) https://digital.nhs.uk/services/data-access-request-service-dars
- Grampian Data Safe Haven (DaSH) https://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven.php; https://www.abdn.ac.uk/toolkit/systems/safe-haven/
- Health and Data Research UK (HDRUK) – including the UK Health Data Research Alliance, the Health Data and Research Innovation Gateway and Research Hubs https://ukhealthdata.org/; https://www.hdruk.ac.uk/infrastructure/the-hubs/
- NHS Digital https://digital.nhs.uk/
- NHS Population Health and the Population Health Management (PHM) Programme https://www.england.nhs.uk/integratedcare/building-blocks/phm/
- Secure Anonymised Information Linkage (SAIL) Databank https://saildatabank.com/
- The Ada Lovelace Institute: Rethinking Data Programme https://www.adalovelaceinstitute.org/changing-the-data-governance-ecosystem-through-narratives-practices-and-regulations/; https://www.adalovelaceinstitute.org/the-foundations-of-fairness-for-nhs-health-data-sharing/
- The Alan Turing Institute ("The Turing"): Health programme https://www.turing.ac.uk/research/research-programmes/health-and-medical-sciences
- The Connected Health Cities Programme (2016-2020) https://www.connectedhealthcities.org/; Also see: Northern Health Science Alliance (NHSA). (2020, May). Connected Health Cities: Impact Report 2016-2020 – Delivering Trustworthy Data Driven Improvement in Care for our Patient Population. Retrieved from: https://www.chc-impact-report.co.uk/.
- UCL Data Safe Haven https://www.ucl.ac.uk/isd/services/file-storage-sharing/data-safe-haven-dsh
- UK Biobank https://www.ukbiobank.ac.uk/
- Understanding Patient Data https://understandingpatientdata.org.uk/
- Wessex Care Records (WCR) https://www.wessexcarerecords.org.uk/

Appendix B: Ethics checklist for a Social Data Foundation

The following ethics checklist for a Social Data Foundation is derived from a review of ethical principles, including the CARE Principles (2018), the 5 Safes (UK Data Service), the FAIR Principles and the Trust Principles for Digital Repositories (Lin et al., 2020).

In each case, the target actor in the data lifecycle responsible for answering the specific question is shown. Advisors to the data steward would typically review these, and would be expected to have expertise in data protection law, research ethics and data trust policies. Such an Advisory Board would function as an independent panel similar to institutional review boards but with additional knowledge on data protection and trust policies. The Social Data Foundation would ultimately be responsible for acting on their recommendations after they have reviewed the following checklist:

ETHICS CHECKLIST FOR A SOCIAL DATA FOUNDATION: TO EVALUATE PROPOSED SHARED DATA ANALYSIS PROJECTS

1. What are your plans to engage with the participant cohort? *(Target actors: data providers and data users)*
2. How will you accredit those collecting the data? (OPTIONAL) *(Target actors: data providers and data users)*
3. Does the data subject / participant retain control of their data? *(Target actors: data providers, Data Foundation and data users)*

[38] Note that URLs provided are correct at time of writing (November 2020).

4. How can the data subject / participant access their data? *(Target actors: data providers, Data Foundation and data users)*
5. What are your plans to give benefit back to the participant cohort? *(Target actors: data providers, Data Foundation and data users)*
6. Will your research involve data or the participation of those from vulnerable groups? *(Target actors: data providers, Data Foundation and data users)*
7. Will your research involve data or the participation of those in marginalised groups? *(Target actors: data providers, Data Foundation and data users)*
8. What are your plans concerning: (a) collaboration, (b) non-malevolence, (c) beneficence and (d) justice? *(Target actors: data providers, Data Foundation and data users)*
9. How do you publicise your data? *(Target actor: data users)*
10. How do you publicise your research outcomes? *(Target actor: data users)*
11. How would other researchers access your data? *(Target actors: data providers, Data Foundation and data users)*
12. What metadata standards do you support? *(Target actor: data providers)*
13. How would researchers link your data with other datasets? *(Target actors: data providers, Data Foundation and data users)*
14. How have you made sure that your data are complete and consistent? *(Target actors: data providers, Data Foundation and data users)*
15. How will the public be made aware of data and data services associated with them? *(Target actors: data providers, Data Foundation and data users)*
16. What measures are in place to ensure the integrity and quality of the data held by the repository: (a) when collected and (b) in the longer term? *(Target actors: data providers, Data Foundation and data users)*
17. What measures are in place to understand and demonstrate the expectations of communities of user and of practice? *(Target actors: data providers, Data Foundation and data users)*
18. How will you ensure that the data repository continues to be accessible? *(Target actor: Data Foundation)*
19. Describe the mechanisms (including operational processes) to ensure: (a) the reliability of the Data Foundation and (b) the security of the data held by the Data Foundation. *(Target actor: Data Foundation)*
20. What approvals do you have and from whom to use the data? *(Target actors: data users and (where applicable) data providers)*
21. What are the objectives of the research and how will the data be used? *(Target actor: data users)*
22. How will you protect the data you are using? *(Target actors: data providers, Data Foundation and data users)*
23. How will research outcomes be publicised? *(Target actors: data providers, Data Foundation and data users)*
24. What measures are in place to reduce the risk of re-identification? *(Target actors: data providers, Data Foundation and data users)*
25. How will you share outcomes with individuals or communities who provided the original data? *(Target actors: data providers and data users)*

Appendix C: List of acronyms

ACRONYM	GLOSS
ACM	Association for Computing Machinery
AI	Artificial Intelligence
CAG	Confidentiality Advisory Group
CORMSIS	Centre for Operational Research, Management Sciences and Information Systems
CHIA	Care and Health Information Analytics
CHIE	Care and Health Information Exchange
DMZ	Demilitarized Zone
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
ERGO	Ethics and Research Governance Online
EU	European Union
FSMM	Federated Service Management Methodology
GDPR	General Data Protection Regulation
HDRUK	Health Data Research UK

HIOW	Hampshire and the Isle of Wight
HRA	Health Research Authority
ICO	Information Commissioner's Office
ICS	Integrated Care System
IG	Information Governance
IRAS	Integrated Research Application System
ISO	International Standards Organisation
IT	Information Technology
ML	Machine Learning
NHS	National Health Service
ODI	Open Data Institute
OECD	Organisation of Economic Co-operation and Development
RIS	Research and Innovation Services
SAIL	Secure Anonymised Information Linkage
STP	Sustainability and Transformation Partnership
TRE	Trusted Research Environment
TTPI	Trusted Third-Party Intermediary
UHS	University Hospital Southampton
UK	United Kingdom
UKAN	UK Anonymisation Network
UoS	University of Southampton
WSI	Web Science Institute

Appendix D: List of figures

Appendix E: List of tables

Glossary

For the purposes of this white paper, we define the following terms as follows:

Advisory Committee. A group of individuals external to the Social Data Foundation – with a wide range of expertise related to health and social care transformation (e.g. health and social care services, cyber-security, data governance, health data science, ethics, law) – that provides advice to the Social Data Foundation Board on matters related to data sharing.

Anonymised data. The definition of anonymised data is provided by GDPR Recital 26, namely *"information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."* Although strictly speaking, Recital 26 is not binding it has been used by the Court of Justice of the European Union and other national courts to interpret the concept of anonymised data. As a matter of principle, two different processes can lead to anonymised data: a risk-based approach to aggregation (i.e., data is aggregated, e.g. to produce counts, average, sums) or a risk-based approach to de-identification (i.e., data remains at the individual level). In both cases, data and context controls should be combined to guarantee that re-identification risk is remote over time.

Citizen Representative. An expert in data governance, who is mandatory member of the Social Data Foundation Board and oversees the administration of citizen participation and engagement activities to ensure that Social Data Foundation maintains stakeholder approvals. In particular, the Citizen Representative shall (i) contribute to the definition of use cases as well as (ii) create and manage a framework for citizen participation and engagement activities.

Confidential Patient Information. The legal definition of 'confidential patient information' is provided by s.251(11) of the National Health Service Act 2006 as follows: *"[…] patient information is "confidential patient information" where—[/] (a)the identity of the individual in question is ascertainable—[/] (i)from that information, or [/] (ii)from that information and other information which is in the possession of, or is likely to come into the possession of, the person processing that information, and [/] (b)that information was obtained or generated by a person who, in the circumstances, owed an obligation of confidence to that individual."*

Data Analyst. A person who works for a data user and carries out data linking, query and analysis as part of a shared data analysis project.

Data Foundations Framework. A model for good data governance based on existing foundations laws enacted by the Channel Islands – centred on six fundamental components: (i) a comprehensive rulebook; (ii) an independent governance body, (iii) an inclusive decision-making body; (iv) a standardised process for flexible membership; (v) a trust-enhancing technical and organisational infrastructure; and (vi) a well-regulated legal structure.

Data Protection Officer (DPO). A standard role (whose appointment in some instances is mandatory under the GDPR) for organisations who process personal data to oversee the processing to ensure that it is compliant with GDPR obligations and respects data subjects' rights. For the Social Data Foundation, the DPO is responsible for overseeing the processing of any personal data within the Social Data Foundation and advising on compliance with the GDPR and the implementation of controls to address the risk of re-identification when data providers' data is linked in response to data users' queries. The DPO works closely with the Independent Guardian who is responsible for overseeing the processing of all types of data.

Data Provider. An entity who makes available specified health and social care data for (re-)usage by one or more data users as part of the Social Data Foundation. A representative of the data provider could act as a member of the Social Data Foundation Board.

Data User. An entity who discovers, uses and/or re-uses shared data made accessible via the Social Data Foundation. It is important to notes that a data user can also be a data provider. A representative of the data user could act as a member of the Social Data Foundation Board.

De-identified Data. Individual-level data that has been subject to both data and process controls such that the re-identification risk can be considered to be remote. De-identified data should be considered to meet the legal standard for anonymisation.

Dynamic Linking. Bringing together two or more sources of health and social care data on demand according to the specific parameters of a data user's query.

External Auditor: A body independent to the Social Data Foundation who is responsible for auditing or certifying its performance, conformance to standards or compliance to regulations.

Independent Guardian. A team of experts in data governance, who are independent from the Social Data Foundation Board and oversee the administration of the Social Data Foundation to ensure it achieves its purposes in accordance with its rulebook i.e. that all data related activities realise the highest standards of excellence for data governance. In particular, the Independent Guardian shall (i) help set up a risk-based framework for data sharing, (ii) assess the use cases in accordance with this risk-based framework and (iii) audit and monitor day-to-day all data-related activities, including data access, citizen participation and engagement.

Personal Data. The legal definition of 'personal data' is provided by Article 4(1) of the GDPR as follows: *"[…] any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".*

Platform Facilitator. An executing officer, usually supported by a team, who oversees the technical day-to-day operation of the Trustworthy Data Sharing Platform, including the provision of infrastructure and functional services for data providers and data users, the implementation of governance policies, support services for other roles where required, provides an access point for data users and hosts a data repository if required.

Positive Health and Social Care Transformation. The continuous transformation of health and social care services in response to societal demands and advances in clinical practice, medicine, and technology.

Progressive Digitisation. The transformation of large and complex systems from analogue to digital form through incremental steps in different parts of the overall system.

Pseudonymisation. The legal definition of pseudonymisation is provided by Article 4(5) of the General Data Protection Regulation (GDPR) as follows: *"the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such 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".*

Social Data Foundation ("The Data Foundation"). A new data institution for multi-party data sharing between the Council, Hospital, University and other interested organisations to enable positive health and social care transformation. The Data Foundation would act as a trusted third party intermediary (TTPI) for good data governance – based on the data foundations framework – with strong citizen representation.

Social Data Foundation Board. An inclusive decision-making body whose appointed members represent the interests of stakeholders – data providers, data users and citizens – and therefore must include a mandatory Citizen Representative. The principal responsibility of its members is to administer the Social Data Foundation's assets and carry out its objects, including the determination of objectives, scope & guiding principles as well as governance policies & regulations through the comprehensive rulebook. All members shall carry out their duties both lawfully and ethically.

Trustworthy Data Sharing Platform. The trust-enhancing technical and organisational infrastructure provided by the Social Data Foundation, which has potential to offer a range of data hosting and support services, and therefore enable responsible and trustworthy data discoverability, sharing, usage and re-usage.

Trusted Third Party Intermediary (TTPI). A responsible and reliable entity that facilitates data sharing interactions for projects related to health and social care transformation between data users and data providers – whose involvement is acceptable to all parties involved.