UNIVERSITY OF LIVERPOOL

Newcastle University

Centre for Data

ESRC Smart Data Research UK Strategic Advice Team

REPORT 3: INTERVIEW FINDINGS

Rachel Franklin, Jessica Crosby,
Jeanette D'Arcy, Simeon Yates,
Elena Musi, Omar Guerrero, Seth Spielman

October 2023

DMSI

Digital Media and Society Institute

Centre for Urban & Regional Development Studies

CURDS

CONTENTS

INTRODUCTION

This report was commissioned by Smart Data Research UK – an Economic and Social Research Council (ESRC) data infrastructure programme. The purpose of this report is to provide context, input, and early-stage recommendations regarding the overall structure of the Smart Data Research UK Phase 2 Investment. It was prepared by Jessica Crosby at the University of Newcastle and Jeanette D'Arcy at the University of Liverpool, members of the Strategic Advice Team funded by Smart Data Research UK to give independent strategic advice between 2022 and 2024.

Disclaimer

The findings and conclusions presented in this report are solely those of the researchers and do not necessarily reflect the views of Smart Data Research UK or the Economic and Social Research Council.

EXECUTIVE SUMMARY AND HEADLINE RECOMMENDATIONS

This report brings together the results of the interview phase of the Strategic Advice Team's (SAT) work on the Smart Data Research UK programme. Findings from interviews were analysed thematically and five distinct themes emerged:

Infrastructure

A priority for infrastructure is to build flexibility and accessibility into the foundations of the programme, therefore ensuring that whatever infrastructure is adopted can be sustainable for long-term research use in a sector that is constantly innovating. It was stressed by participants that the programme should not seek to 'reinvent the wheel' when it comes to programme design, as there are existing frameworks and tools available on which the Smart Data Research UK programme can be adapted. It would be desirable, also, to include a variety of different projects which could start small and be scaled up, to promote better equity of access amongst the research community. It was highly recommended to provide data on an open access basis. This could be enacted as a condition of funding, not just at the end of a research project, but offered in staged intervals throughout the process. The importance of the programme being able to meet demands of researchers at different career stages was stressed as an aspect of open access, so participants were unanimous in their assertion of the need for a centralized 'hub'[1] to help standardize and collate knowledge, guidance and best practice. A key question remaining for the programme is whether to offer a model where researchers approach the services with a particular type of data or dataset they would like to access, and services try to get them access to that data, or one where researchers approach with a topic or question and services direct them to what is available that could serve their needs.

The programme should:

- account for interoperability, not only in the initial design of the programme, but thinking of how this will enable sustainability of the programme over time.
- approach the building of infrastructure not just with a focus on how data is stored, linked and described, but do so in parallel with a focus on how data will be and actually is used by researchers, and how researchers might want to work with data in the future.
- create a central 'hub' as a 'first stop' for smart data researchers at all stages of careers, which should:
  - provide a trusted source of information and guidance on licensing and legal standards.
  - act as a community of practice and communication.
  - help produce standard guidance for metadata quality and description.
- lead the way in ethical governance, including the production and collation of ethics guidelines and the latest research.

Skills, Training and Outreach

A key recommendation for skills, training, and outreach deals with investment not just in technology, but also in people, particularly relating to the development of interdisciplinary expertise. Researchers in the social sciences identified gaps relating to data science and vice versa. A particular focus was centred on responsible research, ethics, and 'add-ons' relating to specific projects or data sets. The programme has an opportunity to lead on emerging intersections between AI and smart data and should be prepared to advance materials and training as this field develops. Data providers felt that the programme should focus on outreach and the Sci-Comms skills needed to 'tell the story' of smart data.

The programme should:

- create a clear outreach programme aimed not just at industry and academia, but public awareness and education around the benefits of smart data Research.
- focus on three key areas of training and skills:
  - Computational/data science skills for social science researchers
  - Social science (methodological/ethics) skills for data science researchers
  - Sci-Comms and outreach skills
- lead the way in training and expertise relating to AI and smart data research.

Building Relationships

The programme has an opportunity to fill a gap in existing smart data resources around communities of practice and information exchange, as there are few peer-to-peer support networks available between different programmes. Researchers are keen for the programme to provide access to industry and government, and therefore a key focus will be the building of trusted relationships with potentially risk-averse data providers. Building these relationships should promote iterative processes in helping providers to understand what data they have, and what its value is.

The programme should:

- focus on building communities of practice and knowledge exchange; this could be done through a forum or peer-to-peer support structures.
- focus on creating and developing long-term sustainable

1 It should be noted that the ESRC also adopt the term 'hub' as a signifier for the programme team, so this terminology may need to be adapted moving forward to avoid confusion.

Data Sharing Agreements, as well as quality standards and infrastructure that will create sustainable resources.

- create and fund 'liaison' roles designed to bridge the gap between industry and academia.
- closely consider ethical governance of relationships between industry and government.

Legalities, Licensing and Ethics

The SDR UK programme has the opportunity to help centralize and collate ethical guidelines relating to smart data research, as there are not currently agreed-upon standards for this area. The programme should be able to gather a bank of experts available to researchers for advice on applications. Legal standards are generally considered clear, and in some senses are relatively straightforward; ethical standards require a deeper level of interrogation and are currently approached on a person-by-person basis.

The programme should:

- lead the way in focusing on ethical governance, including the production and collation of ethics guidelines and the latest research in this area.
- lead the way in producing and collating standardised documentation relating to legalities and ethics in the smart data field.
- be mindful of legal precedents taking priority over ethical considerations.

Risks and Security

It was again stressed that when looking at risks and security, the SDR UK programme should not attempt to 'reinvent the wheel' and should make use of existing security measures and tools. Use of existing cloud-based resources was strongly recommended, particularly as these resources can be easily updated and augmented to accommodate new innovations in the smart data landscape. Key concerns over reidentification, siloing and data leakage can again be addressed by investing in relationship building, community and public outreach, and ethical expertise.

The programme should:

- pursue public policy research.
- pursue a tiered approach to secure access, and/or create pathways to 'researcher passports'.

INTRODUCTION

The rapid adoption of digital devices has initiated a new era of possibilities for smart data. Smart data is the term used for information generated through interactions with digital devices, including mobile apps, social media, wearables, satnavs, sensors and more. This data is distinguishable from large-scale unstructured data (i.e., Big Data) because it is intelligently processed, meaning that it has made an integral transition from *information* to *knowledge*, rendering smart data a more reliable — and readily utilizable — source for technological development, thus distinguishing smart data as a key resource of the 21st century (Lenk et al, 2015). Smart data has great value to areas such as finance, healthcare, business, security, and smart cities, not only in terms of ongoing processes, but also in anticipation of future opportunities in these spaces. With global attention turning towards advancements in AI, for instance, smart data can figure as an integral resource for refining the analytical models that AI systems monitor. An essential feature of smart data is that it is considered 'actionable knowledge' resulting from the intelligent processing of unstructured big data (Souifi et al, 2021); it can offer analytics of greater quality and speed, allowing AI systems to generate the right kind of predictive insights to produce meaningful impact.

The impact of using smart data for technological innovation is widely celebrated, but what is less talked about is its role in supporting social science research. Smart data is an incredibly rich resource for understanding social life, offering intimate insights into people's everyday engagement with media and technology. Smart data has already been acknowledged as contributing to a growing research interest in the 'quantified self' (Wolf and Kelly in Lee, 2013), that is, the use of new digital devices to obtain automatically collected data about personal activities, which promotes data-enabled self-insight, effectively turning the 'intimate' into the 'informational' (Calvard, 2019). Reciprocally, this also implicates smart data as an important resource for finding the 'intimate' in the 'informational', which is a key remit in social science praxis. The social sciences have long pursued critical contextual understanding of everyday practices through interrogation of cultural norms and institutional problems (Niederer & Chabot, 2015). Smart data can contribute to this field by offering richer understanding of the conditions and consequences of innovation, which includes critical examination of implicit assumptions built into prevailing innovation agendas and practices (Rommetveit et al., 2017). Social science interventions/intersections with smart data present an opportunity for tech practitioners to reflect on elements of usage that may be invisible to them, but that are highly relevant (and potentially problematic) from social and individual perspectives (Rommetveit et al., 2017). A positive impact of social science collaboration is a renewed focus on ethical requirements, which steers researchers and developers alike away from 'checkbox' mentalities to consider the depth and importance of rigorous ethical governance. This has had positive social (as well as intellectual) benefits, as it underlines the importance of seeing innovation in the tech domain as a distributed and/or networked process, reliant on points of shared understanding between tech developers and the public. Overall, social science collaborations with smart data promote the individual's right to participate in the science of the self, championing a more democratic association between user and producer.

Aims and Objectives

Interviews were conducted with the following aims:

1. To map current work in the field of smart data and gather evidence of best practice and challenges.
2. To gather in-depth data on the experiences of experts working in this field of data management and research.
3. To gather in-depth data on the experiences of data providers.
4. To bring together expert advice on the running of data services for research purposes, with focus on the social sciences.

Methods

The SAT conducted 23 online semi-structured interviews, 16 with members of the Advisory Group/other experts and seven with Data Providers. Participants were recruited through existing contacts of the SAT team (e.g., the project expert Advisory Group), as well as those provided by the ESRC programme team and through snowballing. Ethics approval was gained from the University of Liverpool's ethics committee; written and verbal consent was sought from all participants. Data collected was anonymised during the transcription process. Interview transcripts were subjected to thematic analysis and coded using NVivo by two researchers working iteratively to first develop a coding framework, then over email and online meetings to develop and apply this framework to the data.

Please note that in the report to follow, interviewees are referred to as 'P' for 'Participant' and participant numbers have been used to maintain anonymity (e.g., P1, P2 etc.).

FINDINGS

Infrastructure

Across the interviews, infrastructure was a key theme as this is a topic at the heart of the SAT's remit and as such constituted an explicit line of questioning. Participants were likewise often eager to offer their experiences, concerns and advice in this area as one which they considered crucial to the success of the Smart Data Research UK programme. Many interviewees emphasised the importance of 'not reinventing the wheel', pointing out that there has already been good work developed in related ESRC projects and others. The work of the Consumer Data Research Centre (CDRC) and the Urban Big Data Centre (UBDC) were often raised as examples both from participants already affiliated with the two centres, as well as a selection of participants who had no affiliative links, suggesting these centres can act as models of best practice to help guide the SDR UK programme, as well as highlighting existing technologies that are already fit for purpose. Participants stressed that they have already learned many 'painful lessons' (P5) and as such were in a position to offer expert advice, and interviewees felt that the programme was timely and in a good position to take advantage of current best practice:

> *… growing from here is easier than it was to grow up to this point… that whole process [of gaining access to data] is time consuming and painful… once you have it in place, it can scale well. - P6*

Participants often spoke of having started their work in this space with little available to them but were enthusiastic about what technologies are now available that the programme could utilise. Several spoke about the use of cloud computing, and some gave specific examples of solutions like Azure and AWS as 'scalable and it's cheaper and it's updated' (P1); while there are associated costs, these were felt to be worth it, as costs can be kept low and services sustainable:

> *I pay a certain amount every month, it gets automatically deployed for me. It was configured for me. It gets automatically backed up. Security patches are applied… managed services in the Cloud to avoid you having to build things from scratch and to have the ongoing cost… – P11*

As will be touched on elsewhere in this report, while useful frameworks are already available in places, it is worthwhile to 'bake in' flexibility in infrastructure to ensure any potential issues arising can be worked around and will not create impediments to the smooth working of the programme overall.

Flexible design utilising existing technologies, structures and expertise

Participants emphasised the need for flexibility. P11 advised that 'you need a variety of different ways to store the data, but you need one central way which allows you to then navigate and pull everything together'. They described as an example a dual structure, with a data warehouse in which structured data is stored, alongside a BLOB store with less structured data, and linkage provided between the two.

The SDR UK remit is ambitious and hopes to target many different types of data as well as different methodologies and approaches across the 'quant/qual divide'. P11 spoke about 'multi-modality', where:

> *… a very general system might be having to cope with video, it might be having to cope with text, it might be having to cope with audio, it might be having to cope with more structured sort of data from questionnaires. So really there's not particularly a one-size-fits-all solution to that. - P11*

Participants working with social media data were particularly keen to emphasise the shifting nature of this space, and the need to approach infrastructure with flexibility in mind:

> *What is available and what isn't… is something that is fairly fluid in the sense that things change quickly, we witnessed not too long ago that Twitter closed its API and access to data for academics. – P12*

Linkage and interoperability have been a recurring theme across SAT activities, and while participants were concerned with how to do this safely and ethically, there was also confidence that this can be done, and that it is already being done. P2 described how their team take requests from researchers, extract data sufficient for the work in question, and make only that extract available to them via a virtual machine accessed in their own office, or the Safe Pod network, or their own 'safe haven', so that 'if the [SDR UK] programme starts… producing data products that might have some kind of personal identifier or household identifier, then we've got the systems to do safe linkage in existence (P2). P2 also emphasised that, not only does the programme need to utilise existing resources, but value will come from using the data SDR UK gains access to in conjunction with other extant datasets and types. Several participants advocated for ESRC to employ its 'soft power' within the programme to leverage information sharing and collaboration between different data facilities, which could also prove beneficial in relation to the need for common space and a community of practice that was raised in many interviews.

Creation of a central 'hub'

Many participants suggested the idea of a central 'hub', both in terms of a central data storage warehouse, and a more general 'port of call' that would be an initial point of contact for researchers accessing the programme's services. The hub could act as a first stop that any researcher, at any level, undertaking smart data research, could use to access the data in the programme's services, as well as resources in the form of guidelines, templates, case studies and standards, and links to available training. Several commented that there was no one place to point to for researchers starting off on a project, or who are new to this sort of research, and the programme has the opportunity to be this place and provide everything a researcher would need to know from the ground up. This could be a solution to issues raised around uncertainties to do with ethics and legalities that can be a barrier to researchers, as the programme could provide a trusted source of information and guidance that is currently lacking:

> *There's not been one hub that's coordinated this in any way. As an academic, for example, I couldn't point at something and say, well, you need to go here to find out what the current guidelines are on this kind of stuff. It's been very ad hoc work. – P1*

P2 described this as an issue of access:

> *… to be able to go to a data [service] with a research question, access resources on whether data exists that could help answer this, how to get it, what training somebody would need to do, where they could get that training, who might pay for that training… it would be nice for it to be accessible to people like us. - P2*

A central 'hub' could also pave the way for the programme to work towards developing the standardised approaches that participants felt would make a real difference in reducing time-consuming and duplicative processes, as well as developing the kind of communities of practice and communal ethics that will be vital in the smart data space:

> *… having some responsible research training and materials and some standard templates and some standard governance procedures would be the way to do that rather than just hope that everybody will make it up as they go along and do the right thing. – P11*

> *… security stuff can be very difficult and very time consuming to build. There are international standards around doing that, that often involve going through checklists that are hundreds of items long and it can take a long time… it's very useful to have a template that can be repeated… you only have to do that very expensive and time-consuming work once. – P9*

P20 described how this would also be reassuring for data providers, potentially easing relationships and allowing for smoother data sharing procedures: 'I think a nice short precis of why this is GDPR compliant and how, that could be put in front of a lawyer, or a GDPR specialist. I think that would be good'.

The 'hub' was discussed as a way to avoid siloing of different services, and a way to bring what could otherwise be disparate services into harmony with the needs of researchers who will likely want to use data from different sources. Some participants felt that the development of such a 'hub' with standardised methods of access and description would be valuable in itself, as this does not currently exist and could be beneficial both to end users, and to providers:

> *…because there's no industry standard way of doing this… It'd be a challenge to get there because the whole bunch of companies have got their particular way of doing things… this is back to making it easy for companies. If you have got a method… then for a number of organisations you'd be able to go in and say… we'll do that match for you because we've got this standard methodology. – P20*

Similar suggestions of centralised access were made in relation to training opportunities:

> *There is enough going on around the UK for them to upskill themselves. There are courses out there that they find, some of these courses are international… there's nothing centralised… having something like that, like a mini NCRM where a student knows where to go to get the required skills I think would be great. – P1*

Many participants recommended following an observatory model as this would be scalable and efficient, avoid researchers producing their own subsets of data that would cause issues of repeatability, and allow for an agreed set of standards followed across all available datasets. The consensus was that the programme should also lead the way in providing quality (standardised) metadata for their resources, and that this should not be an afterthought, but implemented from the start. One interviewee used an analogy to discuss how useful it would be if the programme could standardise descriptions so that they were not about 'data quality from the producer perspective', but 'indicators of fitness for purpose',

> *…which are different. Sufficiently different that nobody knows how to do this at the moment. It's a bit like if you imagine shopping and you want to buy something, how do I know this stuff that I'm about to buy is fit for the thing I want to cook? You sort of know because you've got a lot of background information about what's in that can… We don't have that for data. And you know, you may use that can for lots of different purposes. So we don't have those sort of general descriptions. – P10*

One participant felt that this is an area in which the UK lags behind its European counterparts somewhat:

> *This approach to standardisation, certainly in things like data linkage, classification, metadata standards … I think that in the UK, we've stalled slightly. We've had plenty of opportunity to do this kind of work, but we haven't seen it solidify in any significant way. – P1*

This suggests two key approaches, though these are not necessarily mutually exclusive:

1. Researchers approach services with a particular type of data/dataset they would like to access; services try to get them access to that data. This would move away from an 'opportunistic' approach to data acquisition, instead identifying and targeting particular types of data/data providers of specific interest to researchers in the programme.
2. Researchers approach services with a research question/topic; services direct them to what is available that could serve their needs. This could still follow the more targeted approach described above, as well as potentially offering use for opportunistically acquired data sets.

These approaches could allow for more equitable access through a focus on what researchers are asking for from the services, rather than the agendas of those running the services themselves. This would also allow for both large and smaller projects to have the same service and same point of contact.

Where researchers are collecting their own data, some participants stated that this should be deposited into the services at the end of the project, and would be a way to ensure adherence to the standards set up: the programme could make it a condition of funding for projects under the

SDR UK remit that they follow all guidelines for metadata/description/ethics/legality, and that any collected datasets are deposited with the relevant service at the project's close. This would also be an enabler for interoperability, and while this is currently difficult to achieve, as P10 points out:

> *...there's a global argument that says, if everyone does this, then we all benefit... but it's a matter of well, I only have enough grant to do this bit of work... that may be somewhere where research councils could start to be a bit more specific... [research councils could ask for] your data to be put in a form which is more interoperable, and in certain cases that may mean producing data to particular standards. In others it may be that this data is suitable to be represented as linked data. – P10*

As one interviewee described, the SDR UK programme has the potential to build infrastructure not just with a focus on how data is acquired, stored, linked, and described, but to do so in parallel with the focus on how data will be and is actually used by researchers, and how researchers might want to work with data in the future:

> *... the best thing to do is to have projects in which the collection of the data and 'putting it in'... under the right sort of security regime and with the right interlinking, goes hand in hand with a range of projects which are going to extract value from that data... You've got people actually trying to use the system, and so you discover any issues to do with security, lack of interlinking, lack of metadata, by having real use cases going along with the data collection. – P11*

Via a central 'hub', the SDR UK programme could also provide a repository of the latest thinking/works on ethics and latest legislation, as well as potentially providing opportunities for research in these areas, if it is decided that this is within the scope of the programme.

The creation of a central hub again suggests four possible approaches:

1. Applications to run the central hub are taken separately to those for other services.
2. Those applying to run one of the services are also asked to indicate if they would wish to run the central hub.
3. It is made a condition of funding for one of the services that they also run the central hub.
4. The hub is run by the ESRC.

Interoperability

Interoperability has been a consistent theme across the work of the SAT for the programme. There are many overlaps between the discussion of interoperability and other themes raised in this report, for example one participant commented on the ways in which linking data and descriptions of data were intertwined:

> *...if you can connect the data together, you also need to connect the ideas together and understand the context of each piece of data... whether we're working with a car manufacturer, to try and find out where, why faults occur. And so, you're linking together information about the model of the car, the worker, the parts, the conditions in the factory at the time, you know, to give you a listing view of what's going on is so important. – P11*

Many participants recommended a central data warehouse which could then be linked with other services; one participant felt that a centralised system could sidestep some of the issues around interoperability:

> *... try to make as centralised an approach as possible. If you can build a single repository, single points of truth for people that everyone accesses in the same way, that mitigates the requirement to spend a lot of time building interoperability into the system. – P7*

However, while a central data warehouse was felt to solve some problems, as P11 pointed out this is not a 'magic bullet' solution:

> *...the other approach is to have a central pot and then what you're doing is negotiating with all of these organisations to allow their data to report on the central pot, and that has problems as well because they may not be allowed to do that for legal reasons or because of licencing reasons and so on. – P11*

P13 was confident that the programme would be 'perfectly capable of making sure that there's no siloed working in the establishment of these things', and that establishing different services:

> *... might not be a problem, might be an opportunity. What kind of foundational data you need access to that is common across the different use case areas. So that's a problem that could be solved by just having a look across, and that again, that's more about linked data rather than IT. - P13*

Along with many other participants, P7 stressed the importance of communities of practice and knowledge exchange, suggesting that whatever structure is chosen, it is the human element and relationships between those running and using services that will be crucial:

> *You need to focus your infrastructure on allowing those communities to develop and to exchange knowledge and information with each other. And that hopefully will prevent the worst silos. – P7*

P3 spoke about how, whichever data service model is chosen, it will be most important that the programme remain agile and flexible, identifying any barriers to interoperability that might iteratively emerge in the system, and that the services are created with the goal of working together as a guiding principle.

In terms of the specifics of linking data, interviewees offered many insights into how researchers are and would like to be using linked data, and how this could be achieved in terms of approaches to matching, sharing agreements, consent and ethical use. P1 described one of their projects which links large survey data with social media data:

> *You do that by asking the respondent to the survey if they would mind giving up their social media handle and data. And they say yes to it... you've got their survey responses, then you can also extract all their social media data and then you can start to see how these two things relate to each other. – P1*

P3 talked about their work with the National Statistics Agency in Scotland, where 'they provide a linkage service for us, which means that we, as little as possible, need to use direct identifiers with this data'.

P4 and P3 offered some insights into planning for interoperability and matching particularly of interest for social science researchers:

> *...we're doing quite a lot of fuzzy matching. For research, it doesn't matter loads if there are some mismatches. We can get away with an error rate in there. But not a huge one, so I think it's understanding what are you gonna link the data on? ... What's an acceptable miss rate? ... if you're looking to make these big linked hubs of data, there'll be lots of people in the subsample of the stuff that's not linked, and I think there's a whole interesting cohort of people in there... it's this missed group that we won't be able to capture. – P4*

In light of these examples offered by interviewees, it not only becomes important to account for interoperability in the initial design of the programme but will be necessary to think about how this will enable sustainability of the programme over time.

Sustainability

A key concern raised by many interviewees was the fast-moving nature of the smart data landscape. In terms of data provision, participants spoke about how changes in legislation or in the operating procedures of companies can mean that research is forced to pivot quickly, and that some research may not be possible in future. Many participants gave the example of working with data from *X* (formerly *Twitter*), which has recently stopped offering its free API, affecting many researchers whose work was based on this access. Many participants discussed the desire for data agreements that offer sustainable, long-term, equitable access.

P4 described how it is difficult, once a written agreement is in place, to deviate from that, which may restrict further use, particularly as current approaches tend to follow a 'use it and lose it' model where data can only be used for a particular amount of time in a particular context, and are then destroyed – because these agreements are easier to 'get over the line'. They question:

> *... is there a way in which you could make it longer term, more accessible but still within the bounds of the law and the data sharing agreement? Think about the other potential uses... is it worth putting in the time now to get something that has more flexibility, than it is to maybe go through the whole process again in two years' time, which is what we've been doing and it's incredibly resource intensive... - P4*

This is particularly relevant in the context of academic funding, which is often fixed- and/or short-term, so data agreements are created in pressured environments where projects must be completed on time, and expediency is therefore a key factor. Similarly, P11 bemoaned the lack of standardised quality metadata in data collection, as creating this is time-consuming and can be labour-intensive, without immediate benefit to the collector, but of real benefit to those accessing the data for future use. P19 described their partnership with Google, which has meant that they have built cloud infrastructure over the last few years, enabling them to 'move very quickly' and 'designed to take us to 20, 50 times the capacity that we're currently running at'.

The SDR UK programme has the opportunity to focus on creating and developing long term, sustainable data sharing agreements as well as quality standards and infrastructure that will create sustainable resources, and perhaps to 'push for data sets that can be made more open and more widely distributed' (DPP2). It is well-placed to bring together data providers, service providers and researchers at every step of these processes to ensure focus on the needs and desirable outcomes of each of these groups in parallel.

Participants also spoke about the frustrations of working within an academic funding system in which funding is, as mentioned above, fixed- and/or short-term. This not only means that by the time data agreements are put in place, a large portion of the time allotted for the project has already passed, but that teams can take months or years building relationships with data providers and creating resources which have great value to many stakeholders in government, academia, industry and civil society, then may face closing down because funding runs out. Some participants described partnering with industry to gain outside funding and make projects sustainable beyond the allocated project time, but this will not be a feasible option for some projects. Others spoke about the skills gaps in this field, with P14 describing how 'it's really hard to recruit and convince good social scientists who are also interested in the technical aspects of digital footprints, retaining them in the sector is going to become even harder...'

In terms of the future of data sharing, one participant saw the current climate as something of a crossroads, where the data landscape could develop in two possible directions:

> *...in 25 years time, social sciences data will be held and owned in one or two places that is managed through a system of contracts but can be accessed by these people, for these purposes, using these tools and that can be done with UK data, international data, whatever. And if that's what we're working towards fine and then everyone starts to take those steps towards that. If we're going in a different direction, if it's gonna be more, 'this is my data and it's [mine] and I wanna [make] a lot of money out of it'. Also, I think fine. But we would take different steps then around access and investing in how long you'd get access to the data for and what does it buy you, what it doesn't buy you. - P4*

There was, however, generally a consensus amongst participants that the future direction of travel in the smart data space would be towards more open access, more sharing of data across national and international boundaries, and a greater understanding of the social good that can come of smart data research. However, as P7 points out, 'it will take us and it will take you guys' (i.e., SDR UK) to move towards this.

Discoverability

A point raised across all aspects of this study is the need for high-quality metadata . FAIR (Findable, Accessible, Interoperable, Reproducible) principles stress the importance of metadata in every factor. For data to be Findable and Reusable requires metadata that are 'richly described with a plurality of accurate and relevant attributes' (Wilkinson et al, 2016). Many participants stressed the importance of metadata for discoverability:

> *... in the spirit of what is intended, can I easily reuse this data in my system? Then the answer is usually no for all sorts of reasons. Part of that is about the discovery of the data... is it actually what I thought it was, does it match what the metadata says? If you're lucky enough to have any metadata, of course. – P10*

Findability and searchability is particularly important for researchers or institutes that may be new to working with smart data, or do not work with it on a regular basis and so may need more support. This may be especially pertinent given the programme's aims in increasing and opening access to different kinds of researchers:

...the main barrier is when we go looking to see whether there's data that's been collected on a particular issue, we are not able to find any if it exists... Make it understandable to people from organisations like ours who are... figuring this out from scratch without institutional support... being available for people to go: 'I have this question. Do you have any data?' – P2

Trying to make use of poorly described data was a common experience; P11 used as an example what they saw as a common scenario, in which a PhD student collects data and then moves on from the institution; when the next researcher looks at the data collected, they are left asking 'is this supposed to be a percentage? Why does this one say 354?'. It is key that data is reusable and searchable for researchers who may be accessing it years after collection:

... designing the collection of data or interlinking so that you're collecting metadata about it as well and making that available, making that searchable is absolutely key. In some fields there are standards for the metadata which makes things easier and even sometimes tooling... whereas in other areas there aren't standards and so you've either got to try and come up with one, which can often take many years, or you have to just do your best and make sure it's all documented somewhere. – P11

As suggested in sections above, a key service that researchers would like to see from the SDR-UK programme is a standardised approach to the creation of high-quality metadata, and the programme has the potential to lead the way in this area.

Disparities in Access

Disparities in access can refer to inequity in what data is available, to whom, and whether that data is representative. Imbalances in the way that datasets are collected, developed and described has been linked to social disparities, as data is applied in ways that leave certain populations or groups behind (Ibrahim et al, 2021). P2 discussed the importance of careful methodological focus on equitable practices:

... at local authority level, trying to match data about individuals across different data sets is often done by residents... that works very well if you're a nice middle-class family who own their own home and don't move house very much. But if you are a precariously employed worker who moves house every nine months ... anything linked on residents is going to be less accurate and out of date for you... being thoughtful about what you're losing when you're linking things - because you're throwing out data that can't be linked to, you're lumping it together in unmatched or other [categories] - I'm thinking about the disparate impact of that kind of methodology. – P2

P4 had similar concerns about what populations may be missing from datasets or linkages:

Certainly, it's something we're looking [at in] health and education data and there's not a huge amount, but there's a couple of 100,000 who are not matched. So, who are they?... Potentially they could all be people in private schools who have private healthcare, which is fine. Lovely. They're doing great. It's more likely that they're not in school or they're moving around or they're cycling out of legal migration status... they're not going to the hospital. They're not checking in for medical appointments. So, actually, it's a missed group that we won't be able to capture. – P4

As well as concerns over the equitable collection, linkage and description of data, participants of course discussed inequities of access in terms of what data is available to researchers. Several interviewees raised the issue of how the need to ensure that data is secure can result in issues of inequitable infrastructure, as a small, privileged group become the only ones who can gain access because of their pre-existing relationships with industry or institutions, which allow them 'trusted' status. This can become a self-perpetuating system as described by P5, in which access to confidential and high-quality data affects the careers of researchers, allowing publications in higher-tier journals, larger impact and the ability to do work that is of national and international interest, thus more firmly excluding those who do not have such privilege in the first place.

P3 described a small step towards more equitable access. Developments in security have increased the number of researchers who can use their datasets, as this has shifted from geographically specific 'safe havens' which were typically located in large cities like London or Edinburgh, to more choice in spaces to access datasets securely, e.g., via personal computer or 'safe pods'. Several participants also raised the importance of *longevity* of access (P4, e.g.), as many agreements in the current climate are time limited.

Ideally, researchers would be able to access datasets provided by the SDR UK programme securely on their own devices and will need to consider how to balance security of access with providing that access to as many people as possible. Several participants suggested a 'tiered' approach to security, with different access levels requiring different procedures for different datasets. Another suggestion was to follow the European example of developing 'researcher passports' for those who have experience/training and can therefore be allowed high-level access without burdensome protocols (however, this option would need careful consideration in terms of how to make qualifying for such a 'passport' accessible in itself).

Many interviewees described difficulties in obtaining access to datasets, and how they found solutions in some cases, but these tended to be one-off, *ad-hoc* solutions to specific agreements in specific contexts. As P11 describes:

it's easier to say 'make things open' than it is to do it in practice... industry often has internal data that it doesn't want to release to the world for commercial confidence reasons... what you can sometimes do is to provide ways for industry to upload their data point to a secure store next to the open store and run computations across both of them. So that sort of hybrid approach is sometimes possible, but these are messy problems with no silver bullet, I'm afraid. – P11

As above, participants felt that the key was to have flexibility in design and infrastructure, allowing leeway to those working within data services to be agile, adapt and come up with solutions – but that these solutions can then be held for future use. P14 focused on the kinds of data products that researchers would want access to, again emphasising that flexibility and plurality could be the way forward, especially for a programme which looks to enable access for both extremely experienced smart data researchers, and those new to this area:

I think having access to... some kind of data product while at the same time with some degree of pre-processing and cleaning done can be extremely useful and empowering for most use cases. But of course, in some cases you really do need to have access to raw data... having pathways to be able to access both, with one obviously being easier than the other is probably the preferred way. - P14

Similarly, participants wished for flexibility in the choice of coding languages available to use (P4). Another aspect of encouraging more pluralistic approaches was the question of a 'disciplinary divide' between quantitative and qualitative approaches. While we may wish to move away from such distinctions (Bastow, Dunleavy and Tinkler, 2014), in the current climate many of our interviewees felt that this was still a key issue in this field. P8 saw this as pertinent in terms of the very governance of open practices such as those under discussion here:

... a much more pluralist approach to the governance of open knowledge practices would be really important in getting all parts of academia to engage and feel confident that they're not gonna lose their own knowledge by stepping into that space. – P8

Skills, training and outreach

When asked what capacity building would be needed for the SDR UK programme, many participants identified a need, not just for technological infrastructure, but for investment in 'people as infrastructure'. In particular, many discussed the need for training, skills and recruitment that focused on areas that will be necessary for the programme's success and for future work in the area of smart data:

With many international collaborations the data is not necessarily the key, the key aspect is accessing expertise... there could be people who [are] specialised... but everyone has to have at least a broad understanding of the key concepts in data management and ethics and social processes and computing and open science and so on... – P12

While there is now a growing number of bodies working with smart data, it is a relatively new area and as such participants discussed the importance of creating a community of best practice and knowledge sharing through networks and skills training that will not only develop the field in useful, ethically sound directions, but also create sustainable infrastructure:

Whenever there's new data available there needs to be communities who are working with the data to share and grow knowledge more quickly... there are some very capable people, but they tend to be in small numbers. So, a lot of your knowledge investment is held by a very small number of people, so if they move on, retire, move away, then that's a huge brain drain really. It's about identifying what the skills are and then getting those out to as many people as possible so that you're broadening your skills base as much as you can. – P4

P11 advocated for specific, focused add-ons to existing data science education that could be tailored to particular infrastructure in order to support specific projects or work with specific datasets, 'with plenty of examples so that people can get a warm feeling because they could try something out on the infrastructure to get some results back that they understand'. This interviewee also emphasised the importance of ethics and 'responsible research', echoing comments elsewhere in this report. They described the need for approaches that not only ask for credentials at the start of a project but require ongoing attention that addresses issues as they arise.

Skills gaps and the quantitative/qualitative 'divide'

A recurring concern in interviews was a perceived 'disciplinary divide' and some participants saw investment in training and knowledge exchange as a way to encourage greater interdisciplinary working, and to bring disciplines closer together:

... in terms of them feeling part of a culture of computational social research..., what they see is that quantitative methods get strong investment, discipline-grounded community development often supported by scholarly associations like British Sociological Association, that provides a frame. But for digital sociology or computational social research, that kind of collective culture, I think it's still weaker, institutionally speaking. And I think... we need our computational researchers to be good social scientists – P8

Perhaps unsurprisingly, participants raised skills gaps between disciplines, whereby social scientists may lack the data science or computational skills to manipulate data at an advanced level, and computational or data scientists may lack the methodological skills and background in ethics to enable best practice in sociological work:

That's something that we really need to support and foster and broaden that often requires different models, like peer-to-peer learning... there's some skills that are more invisible skills. To have a critical methodological awareness. What's a good social research design? ... Two different methodological traditions. If you're visualising data that is both web-based and interview-based, what are the styles of visualisation that are appropriate? The awareness of the gaps is much lower when it comes to those kind of skills also ... we've grown accustomed to it being okay to not have those skills. – P8

One key area that was raised in particular by data providers, however, was a gap in skills relating to public engagement and the ability to communicate clearly the benefits of both data sharing and smart data research more generally, both in terms of how to persuade industry to share data with researchers and in terms of how to improve public understanding about how and why data can and perhaps should be shared.

In terms of skills that will be required moving forward, P24 spoke extensively about the advance of AI and how the UK is in some ways unprepared or lagging behind, e.g., the US in terms of expertise and knowledge of this field. They observed that while students appear to be autodidactically gaining some skills in the use and study of AI technologies, there is little in the way of expertise or training that they can be pointed to, or indeed that academics themselves can undertake. Several participants spoke about how the rise of AI technologies could affect open access approaches, as data is critical for training AI processes and may become a source of further tightening of access. Some, however, were also concerned about potentially missing opportunities presented by AI because knowledge in this area is so limited:

I 100% know there will be some sort of AI mechanism out there that could help us code those, but we have no way to access [it]... I feel like in the UK we're just so behind... if we really want to be on the front end, we need to start thinking about how do we use these new tools in this new world, and how do we train people? – P24

Outreach

Just as participants recognised a gap in skillsets between computational science, data science and social sciences, data providers pointed out a gap in the area of what could be considered Sci-Comms (Science Communication – the practice of raising public (as well as academic) awareness of science related topics) (Gerber, 2020) in order to 'tell the story' of the data in a way that is both understandable to the general public, and promotes the good that comes out of social research using smart data:

> *… if you get some very good digital analytics and data engineering resource into there, it's often not their skill set to go out and necessarily do that storytelling piece… you need somebody who can tell the story… The data industry is shockingly [bad] at pointing out the good it does… If I could fix one thing, it would be the data industry being far more front foot about the benefits it actually brings to society. – P20*

Research suggests that public trust is a key factor in enabling the UK to embrace the opportunities presented by smart data (DBT, 2023; CDEI, 2020). Similar to P20, P7 located this issue as one which needs to be addressed via educating the public:

> *… there's an education piece that across the board we need to do … that's the piece that's missing, it's that mass awareness to say this stuff isn't radioactive… Yes, there are risks in misuse of the data, but actually, here are all of those benefits. And if you can get to that point, it will make it much easier for organisations like ourselves to share data with organisations like yourselves. - P7*

The SDR UK programme will need a focus on 'telling the story' of the social good that data sharing and smart data research can do, via a building of capacity in Sci-Comms and social research skillsets.

Building relationships

Relationships between academia and industry

As a key focus of the SDR UK programme will be on securing access to datasets, the question of how best to foster successful and productive relationships between industry and academia was an important theme. Many participants, researchers and providers both, spoke about how risk-averse industry can be, and identified this as a key barrier to access. Of course, the SDR UK programme will need to ensure that they can demonstrate to providers that they are creating secure environments where risk is minimised in order to reassure them, but the consensus was that security is something that people already know how to provide and would not be a barrier *per se*. Rather, participants leaned towards discussion of other more nuanced types of risk such as reputational or competitive, which can be more difficult to address and can raise ethical considerations. P17 thought that 'proper quality security' was 'understood and done', but that relationships could be challenging because:

> *… there are very few people in industry who've worked in academia, and hardly any more people in academia who've worked in senior roles in the industry… the level of misunderstanding of each other's worlds is just colossal. Being able to sit down together and appreciate what's important I think is the key to making this a success. – P17*

P17 here describes the challenge raised by several participants of bridging disparate communities that currently have little contact with each other. P9 suggested that the programme could find ways in which to create bridges between such communities:

> *Finding ways to create a kind of porous barrier between researchers and researchers in the Academy, and researchers and industry… might make data sharing easier because then the organisation isn't giving away the data. The person working on it might be internal. – P9*

One possibility suggested was to create 'liaison' roles that are specifically tasked with this bridging.

P21 spoke about how the process of building more long-term relationships was valuable for both providers and researchers, describing how starting small and scaling up once processes are in place is a viable option for building trust:

> *I would say the first and most difficult [thing] is actually obtaining the data and convincing [providers] to share data with us in the first place… then… going through the process of actually ticking all the boxes on signing the DPA, getting the data sharing agreement nailed down and then looking at the physical security to move the data into our site… quite often [providers] are doing this kind of thing for the first time… they often don't end up knowing all the different bits that need to come together. – P21*

This describes an iterative process, not just of putting in place viable DPAs and legalities, but of discovering what data exists, is useful, and in what ways, for both the provider and the researchers. While this may be time-consuming in the beginning, starting with the funding of some small projects that build relationships on equitable ground and then get scaled up could be an approach that is beneficial for the SDR UK programme.

While providing 'clean-up' and analytics/insight for data providers could be an incentive for owners to share data with centres, one data provider spoke about how providers may not be aware of what data they have, or its value in terms of analytics/insights:

> *… the social data is still a little patchier in terms of their own understanding of what they've got… the digital analytics skills are still in relatively short supply within that sort of area and only about a tenth of the analytics teams would be digital specialists. They have far less natural understanding of the data and the data is far messier as well. – P20*

This understanding is something that the SDR UK services could look to offer as part of the iterative building of trusted relationships described above, that would bring value to the providers as an incentive to share and value to the researchers as it could mean wider access. This sort of service may seem obvious to researchers and/or those running the services but did not seem obvious to the data providers so this is an area where outreach will be important.

P19 went so far as to say that 'getting access to the data is not the problem' at all, rather the issue is:

> *…actually understanding what the real questions of your stakeholders are… one of the benefits of us partnering with academia at the moment is […] because they're working with the transport teams in local government… they really understand what are the day-to-day things those guys are trying to solve. – P19*

This research centre has successfully built a rapport with data providers in order to understand the key questions they want to address, as well as those that are a priority to the research team.

P19 was clear that there needs to be a sense that the programme is providing value and is committed to a relationship with industrial partners in return for access:

> *… you need to have some commitment to this, if the company is seeing there's a bit of a pain in the neck and that you just really want the data then it's not gonna work. – P19*

However, while this might suggest a kind of 'quid pro quo', the programme will need to consider ethical issues and what is acceptable in terms of agreements, as some data providers might wish to place restrictions on research activities in order to protect their interests. P18 described the limits placed on academic research in a relationship with a data centre for which they sponsored PhD research and provided access to data:

> *We would have a form of nondisclosure agreement and an undertaking in the final thesis to just ask for things to be sort of. Redacted. Or for the thesis to be not published for five years or something like that. – P18*

However, P18 also described incentives to work with academia in that it provides the opportunity for research they may otherwise not be able to fund, as well as developing talent:

> *With researchers, I think it is twofold, one is it's nice to sponsor and develop new talent and there is an altruistic side to that, but also a selfish side to that in that we need to bring people into the industry and maybe they'll come and work at [our company]. The other side to it is again quite practical in that within an organisation you always want to do research, but there's actually quite a limited time opportunity for you to do research within a business. – P18*

A challenge for the programme will be 'walking the line' of offering value to industry and building trusted relationships with them, whilst maintaining ethical governance and control over research outputs.

Relationships between academia and government

Similarly, there were ethical concerns raised in interviews around how the programme could work with government and government data. P9 described in detail the challenges they see in terms of the kinds of data available as there are 'lots and lots of people rushing into this space', and more and more data available, but 'nobody will tell you where it comes from. It's usually slipped up by people playing Candy Crush or whatever, looking at the weather on their phone.' These are contrasted with 'other sources of information that are really, really high quality, but really expensive'. This participant sees the programme as an opportunity to provide a space for government bodies to work out ethical ways of working with data:

> *…using academic organisations as a trusted partner to prototype, to help figure out how these data can be used to create positive public impact in a way that is ethical and safe… in advance of governments trying to do it themselves because they feel compelled to because the other sources of information that they've traditionally relied on are much more difficult to collect now. – P9*

This suggests the programme as a kind of 'proving ground' in which methodologies of working ethically with data for social impact can be tried, tested and discovered in order to create ethical standards/norms for the future in a way that utilises existing expertise in these areas to mitigate risk for government bodies.

Similarly, P3 suggested an area of development between government and the SDR UK programme would be developing an understanding of knowledge and information gaps and bringing together data/research projects that would help to fill those gaps in a policy-focused way. For their research, the programme could potentially provide additional data streams that would strengthen and deepen their work.

Similar to P20's comment above that retailers are sometimes unaware of what data they have, or its value in terms of analytics/research outputs, P22 identifies this as an area that can be of benefit when working with government bodies, but requires a longer-term view of this relationship where research questions can be built iteratively in order to be of value to both parties:

> *… it's an iterative process to make sure that they get value and impact from it… it's an interesting challenge where you can have the data but then making credible impactful work out of it is still something that will take a little bit of iteration. – P22*

One participant felt strongly that the programme needs a sharp focus on specific outcomes when dealing with government, and a clear overview of where it will sit in the wider research landscape as well as more generally:

> *… about the interaction with government on this level, whether it's actually aimed or useful at all… just keep that as tightly scoped as possible so you're not doing everything. That is a completely different approach to many research programmes [where] you have a wide breadth of things and see what happens. You might get better traction… if you try and define what's going to happen… – P13*

Legalities, licensing and ethics

Legalities and licensing emerged as a clear concern amongst interview participants. It was asserted by several interviewees that minutiae in legal procedures relating to aspects of security or interoperability were usually some of the trickiest obstacles to overcome, particularly in instances where international collaboration was involved, as end users would often be faced with additional difficulties navigating data licensing practices between different territories. Between data providers and data users, legalities and licensing often became the point at which priorities would diverge, as it proved difficult to balance out industrial measures for compliance (usually relating to risks around security) with researchers' more intellectual attention to ethical practices. It was agreed amongst participants that whilst there are commonly used licensing procedures that can lay a solid *legal* foundation for new data services, there is still not much in the way of a consensus on *ethical* guidelines for approaching smart data, particularly given the nebulous nature of current digital data practices. It was again suggested that the best way to offer more rigorous grounding for both legalities and ethics was to structure the data services in such a way that a central 'repository' (i.e., hub) was positioned to collate legal documentation and allow researchers to communicate and share information with one another regarding ethical practices. Participants again stressed how *ad-hoc* practices have been up to this point in time, and that end users were waiting for a service that could co-ordinate expertise and offer a point of centralized contact and information.

Smart data and ethical greyness

Questions of ethics bring up further points about 'fair practice' in relation to digital trace data, as there was a repeated consensus amongst participants that the type of data handled by the SDR UK programme requires extra foresight to do so in an ethical manner. This is, in one participant's opinion, because smart data lies in an ethical grey spot, as it is data collected passively from consumers, who often do not fully understand the extent to which their information is being harvested by industry and academia alike:

> *… what I mean by digital footprints data are data that you can collect passively from people through transactions or things that they do in the course of their daily lives. I think that data is often - it's ethically very grey. It's often collected without clear consent. - P9*

The interviewee went on to break down the potential divide in priorities when it came to ethical compliance between industry and academia. Whilst participants did acknowledge that researchers also work in ethical grey areas, it was detailed how the pipeline of industrial legal compliance usually begins and ends at user consent:

> *Yeah, I think it's a grey area [ethically]. I mean clearly these companies are doing what they have to do to be legal in the various jurisdictions that they're operating in. Largely that's about consent. Largely that's about being transparent to the users. But I don't think it's enough to have in your terms of service that this data has been collected, and it's not quite enough to actually understand that just because you're using a social media app that has a map of where all your friends are, that that data is then gonna find its way to various other commercial bodies or potentially ultimately into the hands of research… you're never gonna be aware of that. - P7*

Similar issues of public trust and accountability were brought up by researchers, though again, these tended more towards ethical issues of data usage, rather than legal processes. Another participant used an example from their own social media research, for instance, to demonstrate the difficulties of applying existing ethical practices to smart data sources. The interviewee detailed how context collapse is a key definer in understanding how an online user's agreement with the basic terms and conditions of data usage may not extend to their data being used in research, or being otherwise cast further into the public eye:

> *… if you're publishing as a person with a very small follower number… you expect only a small group of people to be viewing your posts, and therefore that may influence the way in which you're communicating. You may engage in posts that are much more personal or hateful or whatever it might be because you think your audience is so much smaller. You have less of an expectation that a social researcher will be scouring your social media for [a] research project…There is such a thing as a right to be forgotten and all of a sudden you've got a post in [a] journal article that's forever. - P1*

An important concern reiterated amongst both parties, therefore, was that there is often confusion between what is legally agreed upon and what is ethically agreed upon, two subjects which are often (and perhaps wilfully) conflated in this domain:

> *We rely on the domain experts who understand what's allowed in that area, but also the ethics, because as you know, there's a difference between the legal and the ethics. And it's very important not to get those two things confused. - P11*

A core concern with using data collected through data services, then, hinges around legal compliance and ethical conduct being treated as being mutually exclusive, when ethical conduct actually requires more of a vested interrogation of data processes. Participants highlighted how these practices will often come down to instances of personal judgement, again stressing that there is no agreed upon framework for dealing with the ethics of reproducing smart data:

> *I think the ethics stuff is… quite patchy…there's no hard and fast rule here, all they have to do is make sure they're legally compliant. So, they're not falling foul of, say, terms and conditions, but terms and conditions don't always map onto what's ethical or not within our disciplines. Ethics… is still in flux…And there's no right answer. You just have to use your judgement. - P1*

Without a centralising body producing agreed-upon standards, researchers and providers alike are left in the position of relying upon best judgement, which cannot provide a solid or sustainable foundation for best practice moving forward.

Best practice

As identified throughout this report, a repeated theme of interviews is that there is no agreed upon 'best practice' relating to ethics and legalities. The infrastructure and information that is in place is very *ad-hoc* and scattered, and though working from frameworks used by other data services is undoubtedly key to ensuring more centrality in the sector, it cannot necessarily account for new challenges prompted by changing technology and data generating processes. A repeated refrain from interview participants was that ethical considerations should therefore be factored in at the beginning of the 'data journey', which requires consideration from both those managing, and those using, data made available through these services:

> *… we don't really know what best practice is. We know what most people do, which is not necessarily the same… we all have to face ethics committees and so on. But again, we should have processes in place where that really embeds those sorts of things right from the start and it's not just, you know, can I survive the ethics board? It should be, how can this actually make my research better? - P10*

As touched on in the previous section, interviewees were critical of instances where legal precedents could be seen to take priority over ethical factors in terms of data access and interoperability. The concerns voiced here seem to speak to a wider issue across the sector, wherein smart data seems only to be subject to the most cursory forms of accountability (e.g., consent, legalities), thus figuring ethical concerns as 'extraneous' or otherwise 'unnecessary' work:

> *The legal frameworks in which that data is accessed - essentially, that will give you the standards. I'm assuming that most things will go under the Digital Economy Act, in which case it's pretty clear what your responsibilities are in law. And maybe that's the distinction, right? That there's absolute base standards, which are your legal requirements. And then there are slightly elevated standards, which are your ethical requirements. - P4*

Again, and as stressed in other sections on disciplinary divides, this could be countered by more qualitative interventions in the 'data journey'. Invoking expertise on undertaking reflexive or otherwise open research design processes would ultimately prove useful in a sector which will continue to innovate, and therefore continue to require iterative ethical attention. It was thus emphasised that skills and training would be a crucial element in making sure that ethical expertise was 'baked in' to the programme offer, ensuring that the programme is able to extend the right expertise, access, and education to appropriately account for ethical data design:

> *There's a lot of interest in responsible research…there's a set of ways in which you are expected to unpack your ideas for what you're going to do and think about the consequences. The idea is that everybody should be educated at the beginning of the project, but you should also be aware as you go along because if… suddenly new results are coming up and somebody says oh, I realise now we could do this, you need to think about well, if we did that, is that a reasonable thing to do? Is it ethical?… having some responsible research training and materials and some standard templates and some standard governance procedures would be the way to do that, rather than just hope that everybody will make it up as they go along and do the right thing. - P11*

International collaboration

As expressed in previous sections, issues relating to licensing or legal standards were often revealed (or exacerbated) when discussing international collaboration between data users. Though the SDR UK programme will be created with a UK context in mind, key lessons can nonetheless be learned in understanding where points of friction between international collaborators emerge. In reviewing participants' comments, it became apparent that tensions between what is *legally* viable, and what is *ethically* viable, were again identified as possible hindrances to productive relations. Where one was able to secure legal permissions, for instance, this would not necessarily transfer to ethical permissions, as one interviewee expands upon in the context of their own research:

> *It went through all the procedures for ethical approval and so on. But then at the stage of publication, for some journals in public health, something that was required was not only ethical approval from the country where the project originated, but also ethical approval from each of the countries where the data was collected. This was all within Europe and the US… it's one example where it's important to have contacts with different institutions. - P12*

Establishing lines of communication between research institutions, at the end user's level, and data services, at a provider's level, was reiterated across these interviews. The bureaucracy involved in attaining both legal and ethical permissions led one participant to describe the process as resembling a state of 'inertia' (P14), a state which, it was suggested by one UK-based participant, had been further complicated by vague post-Brexit guidelines and lost funding opportunities within the EU. In these instances, distinctions emerged once again around legal frameworks working as proprietary groundwork for research:

> *GDPR… has been the most specific set of requirements in any jurisdiction. There are similar legislative frameworks afoot in Canada, in California, but the GDPR is probably the most…I won't say strict, but most 'detailed' set of requirements that we need to follow. So that tends to set the bar that we then follow everywhere else. - P7*

Issues in international collaboration don't necessarily stop at international legislation, then, particularly as there are similar legislative frameworks afoot in other countries. The main issue comes from what one interviewee identified as 'order demands' (P10), this being the specific minutiae of research design that leads a project away from established guidelines. In this regard, participant attention turned back towards the concept of the data 'hub' and what this might be able to afford researchers in terms of legislative and legal expertise:

> *[We need] some kind of repository in terms of what the legislation looks like across the planet and how that might enable international collaborations. - P1*

> *We completed a project which should be published fairly soon, which is around establishing good legal practice, good data protection practice around enabling data to be shared for research in the public interest. The lesson to take from that is for the ESRC to let these [services] share information, share that kind of legal documentation. - P22*

Creating a repository of information (or perhaps potential case studies) could help mitigate some of the bureaucratic strains related to licensing and legalities.

Guides and case studies

Arguments for designing the SDR UK data services to form a sort of centralized 'hub' have been reiterated across this report. Participants were essentially unanimous in advocating how a hub model would offer means of direct communication, particularly if the infrastructure of the programme included a forum for researchers to discuss, moderate, and potentially refine their data usage:

> *Having a dedicated decision forum set up to hear a proposal and ask you questions on it and then help you develop the documentation with legal support, essentially to moderate any perceived risk, if you haven't got that then then I don't think you've got the ability to share data with any authority or confidence. - P18*

> *I think that what you ideally have is a common set of practices and a common set of templates for things like submitting an ethics form and common set of governance for that. So, people who understand the implications of it and can say yes or no to a particular study [and] give advice on how to tweak it. - P11*

In terms of recommendations relating to legalities and ethics, it was again identified that 'piecemeal' funding calls for data services had resulted in a somewhat nebulous provision of guidelines and frameworks. P1 suggested that a co-ordinated resource had been long awaited within the research community, one which may be able to collate and centralize necessary expertise:

> *Again, it's the funding. I think it's just the way the funding's come through in 'bits', as opposed to it being a strategic deployment of funding. I think obviously [smart data] will see a change there. There will be the hub, which will have the ethics component to that. So, I think this is what we've been waiting for quite some time. - P1*

Risk and security for a smart data programme

Quickly-changing tech

When discussing the landscape of data services that currently exist and the risks faced, a risk identified was the often 'turbulent' nature of technological innovation, being a sector that is constantly evolving, often within short periods of time. Participants expressed an awareness of the quick turnaround in the technical landscape and how it will likely affect the programme's priorities going forward:

> *You look back and think, well, maybe it's good that [the SDR UK programme]'s happening now, and it didn't happen five years ago because so much has changed. But then where the hell will we be in five years' time?... this is work that will need to be done continually, over time, as this landscape changes. - P1*

Participants acknowledged that whilst the needs of researchers may aspire to some form of standardization or centralization of data services, the technology that smart data derive from will undergo rapid changes over time, meaning that any form of 'equilibrium' attained in this space can only ever be a temporary measure:

> *So, it has maybe come more standardised, but the technology is changing quickly...new things will come up which will disrupt the equilibrium and those will have different standards in different countries... it's a continuous reaching an equilibrium, destroying it, and move to the next one, and so on. - P12*

Again, it seems that prioritization of flexible design and infrastructure is suitable to account not only for the changing nature of smart data itself, but the numerous changes to licensing agreements, policies, and legalities that new contexts of data production, collection and curation will warrant. Participant recommendations for offering a centralized 'hub' as part of the programme offer would, again, afford a useful repository of prior licensing, legal and ethical case studies on which new amendments or practices could be augmented, as well as a centrally located point of communication whereby end users can deliberate changes to the data landscape.

Disciplinary divides

This section touches on the perspectives of participants concerned with disciplinary divides in data research, and how these divides pose risks to widening participation and knowledge production in the sector. Several participants expressed a desire for multidisciplinary working within the data research community. P8 noted that researchers working on each 'side' of the disciplinary divide have learned qualities and skills that would work to augment data services, making it important to have at least a basic understanding of the fundamentals of quantitative and qualitative work:

> *Typically, people who come from a computational background tend to have those [large-scale data analysis] skills, but they may lack skills in terms of understanding all the ethical or legal aspects related to different types of projects and different types of initiatives, and the implications of different studies in these social contexts. And that's something that's important for them to pick up. At the very least, each group must have at least the basic understanding to understand the other group. – P8*

Several participants described how few mixed-method researchers were operating in these spaces, particularly those coming from qualitative social science backgrounds. Two possible explanations were suggested for a lack of qualitative input in data services. Firstly, it was highlighted that existing data services were built to align with existing computational infrastructures, which are typically developed within quantitative disciplinary frameworks. P8 notes that it would require a form of 'custom' data infrastructure to accommodate qualitative data research:

> *I think there are lots of services that are being developed within particular disciplinary research frames. So, say if you want to work with social media data... services are readily available... with natural language processing... You can sort of step into the infrastructures of linguistics. Computational linguistics in that case, but I find that a lot of the social sciences or humanities research often requires - and is actually interested in - much more customised forms of data capture. We don't have ready-made infrastructure in place. - P8*

Secondly, it was suggested that because quantitative data intrinsically leans towards more aggregate forms of information, it was more ethically justifiable to centre quantitative processes, as this ensures that any data collected will be less susceptible to reidentification. In each respect, both perspectives placed a similar emphasis on existing infrastructures for data services being more quantitative 'leaning':

> *Given the ethical issues surrounding reproduction of posts, especially sensitive posts... we don't always have the opportunity to publish [data] verbatim, or even in a form that would obfuscate its origins. So, we tend to rely more on aggregate representations, which tend to lean towards more of a quantitative visualisation and analysis...not necessarily because of a proclivity [for quantitative work], but more to do with restrictions in terms of ethics... - P1*

Considering this discussion, an important critical point was deliberated regarding disciplinary 'hierarchies', and how these may have translated to digital data spaces. P8 discussed the reticence amongst researchers of different paradigms to collaborate, particularly those coming from qualitative backgrounds, who (arguably) face more professional risks in crossing disciplinary 'boundaries':

> *I do think that there's an issue around disciplinary hierarchies and what are the mechanisms institutionally that can offer a counterweight...I think a lot of the trust that is needed to have that more collaborative approach take root requires that the social sciences and humanities have a bit more of a feeling of, OK, they have our back. You know, we're taking this risk of collaboration. – P8*

The same participant went on to elaborate how dividing data research down disciplinary lines is reductive, as it stifles interdisciplinary collaboration, thus stymieing widening participation in the sector. They asserted that both epistemological perspectives add value to data science, but must be treated with similar degrees of rigour by data services and stakeholders:

> *It's [a] very important point that the computational gives us a space where we can bring quantitative and qualitative approaches together... there's value to be had from coming back to the point that there are different kinds of research perspectives and research lenses and I think none of those research lenses are in fact new, they just need to be applied with the same degree of rigour. - P8*

This call for fair assertion of disciplinary 'rigour' in data services will likely continue to return to the forefront of the conversation around data science in the future, as observations made by participants considered the changing face of digital data processing. In relation to the previous section on evolving tech, participant P14 offered an interesting perspective on the connection between evolving data generation practices and an increasing need for mixed-method expertise. They suggested that a focus on ethics – a core focus of qualitative enquiry - will become vital as data generating processes continue to innovate and diversify:

> *Questions of ethics become really important. A lot of quantitative data scientists, social scientists who have been working with quantitative data, especially secondary quantitative data, don't think very hard about questions of ethics. And I think for them, digital trace data now presents an opportunity to make sure that they really do engage with questions of ethics and don't just ask for an exemption to that, partly because of the diversification of data generating processes. - P14*

This participant reiterated that mixed-methods expertise is valuable in data analysis, as digital trace data, specifically, requires a 'human element' to unpack contextual data in an effective way:

> *If anything, digital trace data allow us to blend those two perspectives [i.e., quantitative and qualitative] ever more closely together. Some of the interpretation that you can do with certain kinds of digital trace data, particularly contextual data, they require us to think in - arguably - very qualitative ways in extracting information. For example, you can have very large language models, but at the end of the day, you still need to make sense of the clusters or the categories that are created, and that's still a task where you need a human in the loop. - P14*

This contribution aligns with those of earlier participants who reiterated the need for a 'balance' between quantitative and qualitative expertise in the creation of new data services.

Bias

Multidisciplinary work was not the only area identified as needing more focused critical consideration. A concern raised by several participants was centred around processes of data curation taking place across data services and how to account for bias in the production of smart data. Where questions of bias emerged, there was a consensus noting that the more 'intensive' the level of data curation, the more information will need to be available to demonstrate to end users just how that curation has been carried out. It was suggested by participant P10 that bias in data is 'rarely properly identified... it's always a bit hit-and-miss as to how well it's described, if it's touched on at all'. Two reasons were offered for this common oversight: firstly, it was suggested that the process of accessing the data is usually so lengthy and hard-won that curation bias becomes an issue that researchers are simply willing to ignore. Secondly, researchers lack the skills to properly identify and navigate bias. In this respect, it was suggested that training on understanding data bias would be effective in raising awareness of these issues:

> *So, we're very good at recording basic stuff about data. We're not so good about actually describing, if you like, the 'soul' of the data. I think training and skills would help there in terms of making people better aware of things like: well, this data is almost certainly going to be biased, is that bias something which matters to you, or is it something which, in this case, doesn't really affect how you want to use that data? It's things like that, lots of questions that people should ask but often don't. - P10*

Participants repeatedly stressed the need for more (qualitative/social science) interventions in data education, to account for ethical and social dimensions of data processing that are often overlooked. Bias, in this respect, is suggested as needing a similar level of critical contextualisation and instruction. Acknowledging contexts of both data collection and curation were highlighted as important in identifying potential instances of bias in the data, particularly as these contexts also impact on whether the data offered is fit for purpose:

> *If you have a data set about a forest, it matters whether it's been collected by an ecologist or a lumber company, right? They're going to be interested in different things. They're going to be asking different questions. They're going to be recording different information. Some of that data might overlap, but some of it won't. And that doesn't mean that the ecologist can't use the lumber company's data necessarily, but it's probably not going to be sufficient for whatever they want to do. - P2*

Participant P10 similarly summarised how discrepancies between the ways data producers determine data quality, in comparison to what an end user perceives as quality, can result in the data offered being unsuitable for use, which in turn runs into risks relating to security and disclosure (e.g. making sure that the data offer is not 'excessive', and thus open to being used in an expediential manner). This contextual divide seemingly results in a loss of some vital information, in that the prioritisation of producers' values (and thus, biases) over the data may not be producing a fully realised account of what the data represents:

> *Where data has metadata, then it tends to be around data quality, but the problem with data quality is, it's a quality description from the perspective of the producer, not of the need of the end user. The producer is happy with what they describe as 'good quality', but the end user is looking for whether it's fit for purpose, which is quite a different thing... it can also be that other data. that's needed to give you a proper semantic understanding of what that data actually represents, may be missing, and usually quite often is. - P10*

Reidentification and data 'leakage'

Interviews conducted for this programme covered both end users (researchers) and data providers, and one area where perspectives from both these parties tended to merge was when considering risks around data privacy and public trust. A potential risk that impacted on both these subjects was the possibility of 'reidentification' of individual level data, as described by participant P9:

> *If you build this large collection of individual information, even if that information is stripped of individual identifiers, there is still a lot of risk of what's called a reidentification attack, where people could take that information and use it to deduce a person's identity and lots of information about them. - P9*

The participant goes on to explain how this would need to be a key consideration for the SDR UK programme, as an issue that severely undermines public trust. From a research perspective, the ethical ramifications of reidentification are apparent, particularly given that reidentification can actually occur through linkage of different data sets:

> *Typically, for example, you may be using data that's been anonymised, but then you combine it with another data set and actually, it's now no longer quite so anonymous as it was. So maybe developing tools that could help to indicate areas like that could be useful. - P10*

Whilst data providers and end users share similar concerns over data security, they often differ in terms of their confidence in their own capacity for secure data handling. For instance, a similar risk to reidentification that was identified during interviews was data 'leakage', a security issue involving unauthorised internal transmission of sensitive data to external sources (Montano et al., 2022). Data leakage was identified by a data provider as one of three key risks industry was invested in addressing:

> *I think industry worries about three risks and I think those risks, in terms of their order of importance, are changing over time. So, I think the three risks are security and [data] leakage, competitive advantage, and bad publicity. - P17*

Interviews with data providers offered some insight into how seriously industry takes data breaches and leaks, particularly given how this often has a knock-on effect on public trust. Participants who were themselves data owners were particularly conscious of the level of responsibility required when taking stewardship of public data:

> *Many of the large companies that have customer data take stewardship of that data very, very seriously. They see it as a relationship that only works if people trust them to be responsible stewards of their information and to do things that don't harm them in any way. When those kinds of things happen, they can have pretty dramatic implications for the company and they also bump up against ethical questions for companies, like, people are giving us this information so we can help them do something. Is it fair for us to give it away? And under what circumstances or what kind of consent do we need? - P9*

Security was therefore identified as a crucial factor when mitigating researcher access to industry data, a subject expanded upon in the next section. Participant P7 shared that they felt that industry had, in fact, begun to become overly 'risk averse' when it came to providing access to data, as attitudes towards sharing data are disproportionate in relation to the type of data under discussion:

> *I think policy could be improved by a better understanding of risk. The classic 'what's the worst that could happen?' question, we're not very good at answering. We end up being very risk averse. We end up assuming that the very worst is gonna happen and the impact of that is going to be catastrophic, when in reality that's going to be very seldom the case… A better understanding of the risk associated with particular applications, a policy driven by what actual society will need as opposed to trying to do this in isolation. - P7*

A possible recommendation that was suggested in helping bridge issues of reidentification, leakage and public trust was to align industry and academia through forums of public education; that is, a joint investment in pursuing public policy research, with the central goal of helping educate the public on just what digital trace data actually is and what it is most commonly used for. Though all participants were united in agreement on the potential of smart data to be used for social good, the lines of communication between industry and academia still present a challenge that would need to be navigated to improve lines of access and collaboration.

Siloing

In order to get researchers and data providers working together, issues of data siloing need to be overcome. Siloing occurs when a repository of data is insulated from other stakeholders, either within their own organisation or external to it (Patel, 2019). Siloing as a theme emerged in interviews primarily in relation to data access. This subject once again shed some light on tensions between industry and academia, with researchers expressing frustration that their use of this data is most often under scrutiny, when the bulk of the public's digital trace data still lies with private companies:

> *I think almost 80% of the wealth of digital footprints data that could significantly improve the way the country is run is sitting with private companies. So, we've been doing a lot of work on getting access to private companies, or private company data. - P21*

Interviewees who were data providers had (in relation to the key risks identified in the previous section) been honest about their intention of preventing bad publicity for their companies: whilst their values also aligned with pursuing data for social good, these priorities diverted once again when it came to reputational risk:

> *…if there was a way to have a clear and compelling public benefit for sharing these things… If the research that is done is to show that their systems are racist or biased or whatever, that's not a public good that those organisations would be excited about, right? Even though that might be true. You don't necessarily want to give away information about your services or the things you provide, just to have a news story about how racist or biased they are. Yeah, a kind of social good that benefits everybody, including the corporation. - P9*

Indeed, one participant used the example of the Cambridge Analytica scandal to identify how the fallout of 'bad press' would likely result in a curtail in access to industry data:

> *I feel that there was a lot of enthusiasm about more open, democratic means of access like API-based research at some point ten years ago… A lot of those open democratic modes of accessing new streams of web-based data, like through APIs… they've just disappeared… or they've been dramatically curtailed so that the sources of data become smaller and smaller or fewer and fewer… there were some key social moments like Cambridge Analytica and the fallout from that. Those of course impacted access… I really think that researchers should be involved in the coproduction of socially relevant data sets, many of which are now held by different companies, and I think we're very far away from that at the moment. - P14*

Despite disparities in motive, it was observed by interview participants that academics have always had an 'experimental' relation to industry, and that with the changing state of digital data services and technologies, it may be more important than ever to allow researchers the room to 'prototype' data for inclusion in public policy (P9). Possible solutions for this issue again tended towards recommendations for a centralized hub (particularly one based on an observatory model) that would fortify community infrastructures, thus creating a community of usage that can effectively (and transparently) share information and expertise. It was observed that investing in 'people as infrastructure', rather than only in technological infrastructure, would perhaps be the most effective means of discouraging siloing, by helping develop community relations:

> *That observatory model I think is the appropriate one to take… So do as much as you can to create one access platform that everyone gets to use and then, you know, try to create as much of a community of use, community of access around that so people are up front and say hey, this is the data sets that we're using and this is what we're doing with it, it would be really nice if these things, where they're missing, they're potentially change requests that we would like to see, and be as visible around that as you possibly can. But it's a lot about managing the community rather than managing the data and the software because fundamentally research is about the people. - P7*

RECOMMENDATIONS

The SAT team offers the following recommendations, based on the findings above.

Infrastructure

The SDR UK programme should:

- account for interoperability, not only in the initial design of the programme, but thinking of how this will enable sustainability of the programme over time.
- approach the building of infrastructure not just with a focus on how data is stored, linked and described, but do so in parallel with a focus on how data will be and actually is used by researchers, and how researchers might want to work with data in the future.
- fund a range of projects, including some small-scale that could then be scaled up.
- create a central 'hub' as a 'first stop' for smart data researchers at all stages of careers, which should:
  - provide a trusted source of information and guidance on licensing and legal standards.
  - act as a community of practice and communication.
  - help produce standard guidance for metadata quality and description.
  - lead the way in ethical governance, including the production and collation of ethics guidelines and the latest research.
  - Applications to run the central Hub could be taken separately from the data services.
  - Those applying to run a service could also be asked to indicate if they wish to run the Hub.
  - It could be made a condition of funding for one of the services that they also run the Hub.
  - The hub could be run by the ESRC.

Skills, training, and outreach

The SDR UK programme should:

- create a clear outreach programme aimed not just at industry and academia, but public awareness and education around the benefits of smart data research.
- focus on three key areas of training and skills:
  - Computational/data science skills for social science researchers
  - Social science (methodological/ethics) skills for data science researchers
  - Sci-Comms and outreach skills
- lead the way in training and expertise relating to AI and smart data research.

Building relationships

The SDR UK programme should:

- focus on building communities of practice and knowledge exchange; this could be done through a forum or peer-to-peer support structures.
- focus on creating and developing long-term sustainable Data Sharing Agreements, as well as quality standards and infrastructure that will create sustainable resources.
- create and fund 'liaison' roles designed to bridge the gap between industry and academia.

Legalities, licensing and ethics

The SDR UK programme should:

- lead the way in focusing on ethical governance, including the production and collation of ethics guidelines and the latest research in this area.
- lead the way in producing and collating standardised documentation relating to legalities and ethics in the smart data field.
- be mindful of legal precedents taking priority over ethical considerations.
- encourage and follow FAIR and open access approaches.
- make it a condition of funding for any projects under their remit that guidance for producing metadata and description of datasets is followed, and that any data collected in the course of projects is deposited with the programme's services.

Risk and security

The SDR-UK programme should:

- pursue public policy research.
- pursue a tiered approach to secure access, and/or create pathways to 'researcher passports'.

REFERENCES

Bastow, S., Dunleavy, P. and Tinkler, J. (2014) *The Impact of the Social Sciences*. London: SAGE Publications Ltd

Calvard, T. (2019). Integrating Social Scientific Perspectives on the Quantified Employee Self. *Soc. Sci,* 8(9). https://doi.org/10.3390/socsci8090262

Department for Business & Trade, *Smart Data: Identifying the features of ethical and trustworthy smart data schemes*, July 2023. Smart Data: Identifying the features of ethical and trustworthy Smart Data Schemes (publishing.service.gov.uk)

Centre for Data Ethics and Innovation, *Addressing trust in public sector data use*, July 2020. Addressing trust in public sector data use - GOV.UK (www.gov.uk)

Gerber, A. (2020) *Science Communication Research: An Empirical Field Analysis*. Germany: Edition Innovare

Mikalef. P., and Krogstie, J. (2019) *Investigating the Data Science Skill Gap: An Empirical Analysis*, 2019 IEEE Global Engineering Education Conference (EDUCON). doi: 10.1109/EDUCON.2019.8725066

Herrera Montano, I., García Aranda, J.J., Ramos Diaz, J. et al. (2022) Survey of Techniques on Data Leakage Protection and Methods to address the Insider threat. *Cluster Comput* 25, 4289–4302 https://doi.org/10.1007/s10586-022-03668-2

Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D., & Denniston, A. K. (2021). Health data poverty: an assailable barrier to equitable digital health care. *The Lancet Digital Health*, *3*(4) https://doi.org/10.1016/S2589-7500(20)30317-4

Lee, V. (2013). The Quantified Self (QS) movement and some emerging opportunities for the educational technology field. *Educational Technology,* 53, pp. 39–42.

Lenk, A., Bonorden, L., Hellmanns, A., Roedder, N., Jaehnichen, S. (2015). Towards a Taxonomy of Standards in Smart Data. *2015 IEEE Conference on Big Data*. FZI Research Center for Information Technology. https://www.aifb.kit.edu/images/0/07/Smart_Data_Paper_final.pdf

Niederer, S. and Taudin Chabot, R. (2015). Deconstructing the cloud: Responses to Big Data phenomena from social sciences, humanities and the arts. *Big Data & Society*, 2(2). https://journals.sagepub.com/doi/full/10.1177/2053951715594635

Patel, Jayesh. (2019) Bridging Data Silos Using Big Data Integration. *International Journal of Database Management Systems.* 11. 01-06

Rommetveit, K, Dunajcsik, M. Tanas, A. Silvast, A. and Gunnarsdóttir, K. (2017). The CANDID Primer: Including Social Sciences and Humanities scholarship in the making and use of smart ICT technologies. CANDID (H2020-ICT-35- 2016) D5.4. http://candid.no/progress

Souifi, A., Boulanger, Z.C., Zolghadri, M., Barkallah, M. and Haddar, M. (2021). From Big Data to Smart Data: Application to performance management. *IFAC PapersOnLine,* 54(1), pp. 857–862. https://www.sciencedirect.com/science/article/pii/S2405896321008491?ref=pdf_download&fr=RR-2&rr=834d3d2fb8e271bd

Wilkinson, M., Dumontier, M., Aalbersberg, I. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data 3,* 160018. https://doi.org/10.1038/sdata.2016.18

DMSI
Digital Media and Society Institute

Department of Communication & Media
University of Liverpool
School of the Arts
19 Abercromby Square
Liverpool
L69 7ZG

Newcastle University
Centre for Data

Centre for Urban & Regional Development Studies
CURDS

THE ORIGINAL REDBRICK