Future Data Services
Data Access Theme Paper 2
Data holder perspectives

Author: Felix Ritchie, Senior Strategic Fellow to FDS.

May 2024

Future Data Services is a two-year review by the ESRC into the operations and management of UK data services. It has five themes: data discovery and curation; data access, user support and training; technology; people, organisations and culture; and ethics, public engagement, and impact. This piece is for the **data access** theme.

Papers produced under this theme are reflections on the evidence gained during the review, augmented by practical experience and familiarity with the literature. They are intended to challenge conventional perspectives and propose new ideas or ways of working. They provide the arguments behind the recommendations of this theme.

The views expressed in this paper do not necessarily represent those of ESRC or the Future Data Services Project.

Reference title: FDS (access theme) Paper 2 *Data holder perspectives*

Contents

Summary of recommendations

discussed in this paper

Recommendations for data holders

A2.1 Data holders should make commitments to making data available for research per se, on the understanding that there will be an opportunity to review specific project applications if the data holder requires it

A2.2 Commitments to making the data available should not distinguish between secure facilities with equivalent accreditation unless the data holder can provide evidence of where the accreditation is not equivalent.

A2.3 Commitment to making the data available should specify the (sub)sets of data to be made available, the broad scope of acceptable projects, and the criteria for any sensitive subsets; these should be wide enough to encompass all likely research, and limits on data or scope should need to be justified by the data holders.

Recommendations for data services

A2.4 Data services should support data holders to identify and prepare datasets for release.

A2.5 Data services holders should build processes around data holders as advisors rather than controllers; data services should provide a mechanism for data holders to raise concerns.

Recommendations for UKRI

A2.6 UKRI should develop guidance on assessing risk and identifying sensible subsets of the data (if any) based on EDRU principles.

A2.7 UKRI should support data holders to develop metadata, use conditions, disclosure control guidelines and other materials necessary for applying for access, by offering support and developing appropriate templates and guidance.

A2.8 UKRI should (continue to) provide funding to support the creation of research datasets as a goal in itself.

A2.9 UKRI should explicitly support the development of effective research networks including data holders and researchers, either directly or indirectly through data services.

1. Issue

Application processes for access to confidential research data vary enormously in the UK. Waiting times range from a week to six months. This has a direct impact on research. PhDs are being advised not to carry out certain types of analysis. Researchers working on short-term contracts (for example on government business) are turning down work on the basis that they do not have sufficient time for application processes to complete. Work is also not being carried out because the scope of existing proposals have been so tightly defined that it precludes additional analyses without a full application.

These delays have multiple causes but essentially these come down to two issues: the need for external approval, and the complexity of the application process. The complexity of the process is covered within FDS (Access theme) Paper 1 *Forms and processes*. What needs to be covered in the application process is covered in Paper 3 [title]. In this paper, we deal with the issue of external approval. This relates to the perspectives of data holders, who we assume have the right to decide whether their data is used for research[1].

External approval matters in two areas

- The approval of the data holders to allow the data to be used for research in general
- The approval of the data holders to allow the data to be used for a specific research project

The paper is based on both our experience of working with data holders in the UK and abroad, and on non-attributable interviews with individuals in UK government organisations who have practical experience of decision-making at a senior level. We focus on government as (1) academic data collections typically have onward access explicit in the consent form (2) statutory gateways are generally permissive (ie they allow research access but do not require it); therefore, an active decision on sharing is required by government departments, and this is not an easy decision for them. Note that interviews concentrated on data holders who had been able to influence change in their organisations, and so were able to speak about 'what works' from positions of experience.

2. Overview: why is data sharing hard?

As we have repeatedly written elsewhere (eg Ritchie, 2014; Ritchie, 2016; Green and Ritchie, 2016; Hafner et al 2017), questions about data governance tend to be framed in a 'default-closed' context: nothing will happen unless the outcome is demonstrably safe. This is usually associated with a focus on worst-case scenarios: what *might* go wrong?

This puts enormous power in the hands of those unwilling to take decisions. The default-closed framing, encourages those decision makers to focus on potential risks rather than benefits and to raise spurious and/or unhelpful 'what-if' questions. As decision-makers in the public sector are encouraged by both incentives and processes to focus on avoiding negative consequences (Ritchie, 2016), this generates sclerotic decision-making processes where activity is substituted for delivery (Behavioural Insights, 2022).

Ironically, the default-closed focus puts more pressure on decision-holders. Consider a decision-maker being faced with the same option phrased in two separate ways:

- Default-closed: "This is what we plan to do; do you think it is fair/appropriate/justified? If you don't approve it we will not go ahead." [implicit: we have an opinion but it's your choice]
- Default-open: "This is what we plan to do; do you think it is unfair/inappropriate/unjustified? Unless you can demonstrate why there is a problem, we will go ahead." [implicit: it's our choice but you can have an opinion]

The default-closed position puts responsibility on the decision-maker rather than taking it away.

[1] What to call the organisations that hold data and use statutory or consensual gateways to allow access to research? 'Data controller' has a specific legal interpretation in the UK and so is not generalisable. 'Data owner' has faced the criticism that the 'true' data owner is the data subject and/or respondent. We use 'data holder' as being a neutral statement; we use it in this paper as shorthand for those making (or influencing) decisions on data access within their organisations.

This is a generalisation; many individual decision-makers (including most of those interviewed) are aware of these issues and trying to promote more positive behaviours. Nevertheless, this is the dominant culture within which decisions are made. Hence we now consider how this default-closed culture manifests itself, and how to change to the default position.

3. Data holders' views

Data holders were concerned about three things:

- Is the data service a secure host for the data?
- Will researchers understand the data sufficiently to produce good research?
- What risk might arise from published outputs?

Data holders were not explicitly concerned that the research topic might be embarrassing to the data holder, although it is partially implicit in the second and third concerns. Cost was also not mentioned as a significant factor, perhaps because those interviewed were more focused on the public benefit rather than operational constraints. However, the experience of the ADRN (Administrative Data Research Network, 2013-2018) highlighted how a poor understanding of costs could generate significant friction between potential partners.

We now consider these in more detail.

3.1 Is the data service a secure host?

Data holders request that any secure facility holding their data should have appropriate accreditation, as part of their due diligence. This is a reasonable request. The question is how to fulfil it.

For all the major UK social science secure research facilities (trusted research environments or TREs, also known as safe havens, data enclaves or research data centres) holding government data, this is straightforward. A key element is the Digital Economy Act 2017 (DEA); this does not apply to all data in all facilities but because it applies to some data in most facilities, it is widely used for accreditation. Using the five safes framework:

Project	Proposal reviewed by a panel designed to satisfy DEA provisions for appropriate use of the data, or other process following same principles.
People	Compulsory training for all researchers (Safe Researcher Training or SRT for social science datasets, MRC training for medical data) plus background checks and guidance on using the facility. SRT was designed to meet DEA requirements.
Setting	All TREs hosting data under the DEA undergo an accreditation process overseen by the UK Statistics Authority (UKSA) to provide assurance that the technical facilities are trusted places to host the data
Outputs	All staff checking output receive on-the-job training and external training, the latter designed to satisfy the DEA provisions for appropriate checking procedures.
Data	Residual

As discussed in Briefing Note 1 *What is the point of a TRE,* the UK TREs are designed to hold the most sensitive research data and accredited as such. 'Data' is the residual: the TREs are accredited to hold all reasonable research datasets, but the actual data that they hold (or make available) may depend upon the research project.

Concerns about the accreditation processes notwithstanding (see Paper 4), it could be assumed that data holders can take the security of the TREs as given. In practice the technical accreditation is

unchallenged, but data holders express concerns that their input into the access process is necessary to ensure that the project, people, data and output elements have appropriate safeguards.

3.2 Data, people, outputs and projects

Data

Data holders expressed the concern that some questions may not be answerable on that data, at least not in the way that the researchers hope. This may be because

- the data is not relevant
- there are peculiarities in the data which researchers are no aware of; for example, PAYE tax data comes with many duplicates because of the way that firms report to HMRC

People

Data holders express the view that researchers might not their data in a way that is appropriate: the researchers

- may not have the technical skills to use the data
- may interpret results and publish findings which are not warranted

Accordingly, advice is needed from the data holder the ensure that the researcher is fully aware of the capabilities and limitations of the data.

Outputs

Finally, data holders expressed concern that outputs may be inappropriate

- Results may breach confidentiality because the generic training given to TRE staff does not cover the peculiarities of the data; for example, inadvertently revealing confidential information about individual GPs because there are multiple observations per GP, or highlighting performance of individual schools
- Findings may be cherry-picked or in other ways unfairly misrepresent the data
- Findings may be publicised in ways which cause embarrassment or inconvenience to data holders; for example, a press release on findings which criticises a key policy and which the department has had no sight of or chance to comment on

In addition, data holders have raised concerns with the Office for Statistics Regulation (in the context of the Integrated Data Service, IDS) that a perceived 'requirement to publish' for IDS projects could lead to intermediate or experimental results being picked up by external parties and confused with final project outputs.

Projects

These three points, plus a concern that the data could be used for an inappropriate purpose, is the basis for data holders view that they should have some role in the project approval process. This is not a universal view, but it is widespread. Note that no data holder suggested that they would vet project proposals based on whether the project itself was 'acceptable' (in the sense of being unlikely to criticise current policies or provide negative evidence of performance).

3.3 What is the real concern?

These different concerns all boil down to one issue: what results will come out of the project? This applies to both the input questions ('will researchers be able to use the data competently?') as the outputs questions ('will we have advance sight of embarrassing results?'). There is a negative interest: data holders want to be sure that unusual confidentiality risks are covered, and that they

have reasonable warning of difficult outputs. There is also a positive interest: data holders may be keen to see results in advance to be able to contribute insights, or to correct errors in the analysis.

These are all valid points; but they do not provide a reason for data holders' involvement at the project scrutiny stage. This is a crude safeguard, as it requires the data holder to try to guess in advance whether results will be acceptable. There are alternative ways to tackle them. For example, several data holders require scrutiny of outputs before release for the TRE, mostly it seems to ensure that confidentiality checks have been done correctly, but this mechanism could also be used to allow in-house communication teams to have sight of potentially problematic findings.

Some data holders argue that there is an efficiency argument: letting researchers loose on data for which they are not qualified or which cannot answer the questions uses up resources. This is a valid concern in the widest sense of public welfare; but the burden of this falls mainly on the researcher who is wasting their own resources, as the marginal cost to the data holder or TRE (once the project is approved) is negligible. Moreover, this is an empirical issue: are the numbers of researchers wasting their time sufficiently high that the data holders should be spending time as the gatekeepers approving the methodological quality of all applications? Is there evidence that the scrutiny of data holders significantly reduces the number of project failures? Finally, even if this were the case, there are alternative remedies. For example, several TREs expect researchers to engage in discussion with them about project feasibility before putting in a formal application.

There is a way forward, involving focusing on specific stages of the approval procedures and testing the relevant choices at each stage. The first part is to separate the decision to make data available in principle, and the decision to allow data to be used for a specific project. The second is to review what information is needed, and explore other ways of achieving the same outcome.

4. Solution Part 1: separating decisions

4.1 General data access

Consider a data holder faced with two questions:

- Should [this data] be made available for research?
- If so, which [accredited] data service is a suitable secure host?

These should be straightforward questions to answer. For UK government departments (including devolved governments), there is an expectation that all data is in principle available for research, and for many data sharing is already underway through the Administrative Data Research UK programmes and funding. In terms of a secure facility, all the DEA accredited facilities are secure hosts, by design.

Once data release for research is not be tied to a specific project, then the data holder can concentrate on the problems around making the data available:

- What data is (most) useful?
- What metadata and user guidance is needed?
- What work is needed to get the data into a research-ready state?
- Is this one delivery, repeated, open-ended?
- Which years are available, and how will updates be managed?
- What link fields will be available, and who will manage any linking?
- Should limits be put on the type of research it can be used for?

Without a tie to a specific project, this allows the data holder to develop a strategic approach to data delivery as an investment, rather than a response to a current demand. Obviously, having an initial project to develop data for can provide both focus and experience, but the key is that the initial project is seen as a pilot, at most, and not an end point. An example of this is the LEO dataset developed by the UK Department for Education, combining education, benefit and pay data all linked to the person. LEO was not developed for any specific project but with the dataset as the end point; there was a general consensus that this dataset would prove invaluable in answering a range of policy-relevant questions.

It may be there the data is of different sensitivity: for example, HESA data on gender self-identity generates small and noticeable sub-groups of individuals. A data holder may therefore split the data into subsets; for example, a data holder may suggest

- 'standard dataset': suitable for any TRE user on an accredited project
- 'sensitive dataset': available for all TRE users but will be subject to additional scrutiny and users will need to undergo training in disclosure control for this specific sensitive subset

To prevent 'defence creep' ('my data is more sensitive than your data'), the onus should on the data holder to justify more than one set of data. The EDRU approach provides a default-open perspective for assessing risk[2], but there is little practical guidance in this area; UKRI may need to develop effective guidance.

Some data holders may argue that, without a clear project, there is no legal gateway, but this is to confuse availability of data with use. Creating and depositing the data in preparation for use is consistent with legal frameworks, particularly for UK DEA datasets. The decision as to whether a specific project is a valid use of the data is a separate decision, evaluated by the appropriate panel when an application is made. Clearly if there is no realistic prospect of research use then the argument for depositing a dataset falls away, but then neither is there any reason for creating the dataset in the first place.

When making the data available strategically, data holders should be able to specify the scope of projects. This should match the dataset: for example, there is no point in stating that 'analysis of the educational experience of transgender students' is an allowable project scope, and not including that variable in the data; or including the variable but not allowing explicit research on this topic.

How detailed should that scope be? Is "Analysing labour market characteristics" sufficient? Yes. The specific project applications are checked to ensure that there is a public benefit, so the use of the dataset can be very loosely specified. Instead, data holders should be considering what research is *not* in scope for their dataset. Phrased this way, it is a brave data holder who would like to state the areas of research that are not permissible – and of course be prepared to justify them.

It may that there are certain areas where the data holder might have legitimate concerns over particular uses, or where the data holder feels that there is a need to make the researcher aware of sensitivities in the data; for example, looking at criminal justice interactions of looked-after children in a small geographical area. There is likely to be value in the researcher engaging with the data holder. But this does not mean that the data should not be made available in principle: what it *might* be used for should not be allowed to limit what it *could* be used for. The specific project application is the place to determine this.

[2] See FDS Strategic Positioning Paper, s9 and s11

Acknowledging the principle of use and seeing preparation of the data for (as yet unspecified uses) as an investment can have multiple gains for the data holder.

First, metadata and other information about the dataset, its values, flaws and limitation can be prepared strategically rather than in response to specific requests. As noted above, data holders often are concerned about sensible use of their data; seeing the dataset as an object separate from use allows it to be described more effectively, and engagement with the researchers is now enhancing that investment in knowledge.

Second, there are potential efficiency gains from actively creating a dataset, and determining parameters such as update and release schedules. The data holder remains in control of the process because it is not directly tied to project timetables. The data holder can also determine 'the' dataset, reducing unnecessary duplication in response to repeated project requests.

Third, if the dataset is an 'object' separate from internal data system, this can make it easier for the data holder to reduce information used for linkage (but not valid for research).

Fourth, this can clarify the question of where the dataset is potentially located. If project requests come through three different TREs, does this mean that the data holder needs to consider each data delivery for security implications? A TRE can be evaluated as a potential home for an object based solely on the data and the accreditation of the TRE, and not on the project or people who will be using it.

Finally, creating datasets in response to demand can create unnecessary security risks. Multiple versions of the same dataset can generate a differencing risk, especially if the parameters to create a dataset are known. Reviewing the inherent risk of a dataset is also a time-consuming operation, which is more likely to ineffective with multiple version of the data; when the data is being combined with other dynamic datasets, this may be an impossible task to complete in a reasonable time. An investment by the data holder in producing 'the' dataset reduce both these risks.

Recommendations for data holders

A2.1 Data holders should make commitments to making data available for research per se, on the understanding that there will be an opportunity to review specific project applications if the data holder requires it

A2.2 Commitments to making the data available should not distinguish between secure facilities with equivalent accreditation unless the data holder can provide evidence of where the accreditation is not equivalent.

A2.3 Commitment to making the data available should specify the (sub)sets of data to be made available, the broad scope of acceptable projects, and the criteria for any sensitive subsets; these should be as wide as possible to encompass all likely research, and limits on data or scope should need to be justified by the data holders.

Recommendations for data services

A2.4 Data services should support data holders to identify and prepare datasets for release.

Recommendations for UKRI

A2.6 UKRI should develop guidance on assessing risk and identifying sensible subsets of the data (if any) based on EDRU principles.

A2.7 UKRI should support data holders to develop metadata, use conditions, disclosure control guidelines and other materials necessary for applying for access, by offering support and developing appropriate templates and guidance.

A2.8 UKRI should (continue to) provide funding to support the creation of research datasets as a goal in itself.

A2.9 UKRI should explicitly support the development of effective research networks including data holders and researchers, either directly or indirectly through data services.

4.2 Project-specific data access

Once the data is agreed to be suitable for research and is hosted or in process to be hosted by a suitable data service, the data holder can now focus on the particular issues raised by the project application. At this stage the availability of the data and the suitability of the platform to host it is not in scope.

Does the data holder have the right to approve or reject projects at this stage? Assuming the broad scope of uses of the data is set at the release-in-principle stage, then 'approval' at this stage should be a simple question: is this within the scope of appropriate uses for this data? If so, then there should be no further need for discussion. This takes us to the second part of the solution: what needs to be checked?

5. Solution Part 2: focusing on facilitating research

The current access procedures often ask for information on feasibility, methodology, intended publication targets, impact. As noted in Paper 1 in this series, there is very little value in the approval body collecting most of this information. Is there a useful role for the data holder? As noted above data holders are largely concerned with scope, feasibility, and outputs. We consider each in turn.

5.1 Project scope

When looking at specific projects, the burden of proof should be on showing that the project is outside the scope of the general purpose. Otherwise, the efficiency gains of specifying a broad scope is lost. If the data holder agrees that this is within scope but that there are some aspects that give cause for concern, these need to be articulated and the scope updated, so that (a) the reason why a project is rejected is clear and (b) similar future cases can be reviewed against the updated scope. This is in line with the precedent/exception model which is best practice for access requests (see Paper 3), and tracking scope changes like this allows creeping scope restrictions to be identified.

5.2 Feasibility

Data holders may argue that the data is inappropriate to answer the question. This should have been addressed in the metadata associated with the data deposit, with sections covering what the data can and can't be sensibly used for. If the data holder wants to enforce this, this should be stated in the acceptable project scope. If not, then *caveat emptor* applies.

Data holders may also argue that the researchers plan to use inappropriate techniques. It is difficult to envisage a scenario where a data holder would be able to insist on a particular technique being used, or how this would be enforced.

5.3 Outputs

Data holders express concern over the potential outputs of a project: they may breach confidentiality, they may have an impact on service/policy delivery, or they may embarrass the data holder in some way.

Potential confidentiality breach is not an issue at the project approval stage. It depends on what the researcher produces, which is evaluated in the output-checking processes. Confidentiality guidelines (eg not publishing statistics relating to specific courts, or schools) should be part of the metadata published with the data deposit, so that researchers have an idea of whether the outputs they want to produce will be allowable, but ex ante a project should not be stopped because of what might be produced.

Some data holders have expressed concerns, particularly for government data accessed through IDS, about a perceived 'publication commitment': that all outputs from a project will be published even if only partial, containing mistakes, in draft, or otherwise inappropriate. This confuses lawful release and meaningful release of results. When statistical findings are approved for release from a secure environment, this is generally the point at which they are no longer consider personal information and so they are potentially subject to FoI requests. This does not mean that they must be published openly, merely that confidentiality is no longer a reason for not sharing that information. For civil service data holders worried about other departments inappropriately using or reporting on their data, the relevant requirement is the civil service code of conduct which gives guidelines on notification.

5.4 Can data holders reject projects?

The implication of the previous three subsections is that data holders should be only able to reject individual projects in specific cases with very good reason. This turns them from approvers into advisors. Data access processes can then be designed with this in mind.

This is a better position for data holders: failing to reject implies that the project is within pre-defined scope, and is not a statement on the societal value of the project. It also means that they do

not *have* to offer an opinion, whereas the approval-based system means that data holders must take a view. This encourages more open discussion.

Recommendations for data services

A2.5 Data services holders should build processes around data holders as advisors rather than controllers; data services should provide a mechanism for data holders to raise concerns.

6. Can this work?

Yes. Evidence of twenty years in the UK and abroad repeatedly suggests that when the following happens

- Someone (a *supervisor*) has responsibility for some part of the researcher data processes
- A more involved group (the *managers*) offers to take over management and allow the supervisor to retain oversight and approve or reject by exception, and
- The managers demonstrate appropriate reporting and due diligence to the supervisor

more often than not the supervisor rapidly loses the enthusiasm for a deep involvement and delegates authority to the managers. Sometimes the supervisors lose all interest once it becomes clear that the mangers are competent, but quite often the supervisors want to retain some interest to show their expertise. This is a much healthier relationship. The supervisors do not have to take responsibility (unless they really, really want to). They can offer an opinion when it is useful and meaningful, but they do not have to have an opinion.

Parts of the UK are already exploring these options. Research Data Scotland is moving toward this two-stage approach. OpenSAFELY gets pre-approval from HNHS data holders using a well-understand frame of reference.

The journey of the Australian federal government since 2015 is the pre-eminent example of national change as a result of decisions to reframe data governance. New Zealand provides an example of a more organic shift, over a longer period, to a default-open model. The Netherlands provides a counter example: there does not appear to be an over-arching philosophical perspective, but a long-standing commitment to identifying and meeting user needs as the prime objective has also achieved a world-leading research environment. The UK would do well to emulate these countries.

It is worth noting that researchers interviewed as part of FDS reported increased use of Dutch and Australian data because of better availability. This was also the case in the UK before the opening of the Virtual Microdata Laboratory (the progenitor of the SRS). HM Treasury's 5th Productivity report in 2004 was largely macroeconomic; almost all microeconomic analyses were based on US studies. In the 7th Report two years later the analyses was largely microeconomic and based on VML studies[3]. It is concerning that we are at risk of returning to the position in 2004.

[3] Ritchie, F. (2008). Secure access to confidential microdata: Four years of the virtual microdata laboratory. Economic and Labour Market Review, 2(5), 29-34. https://doi.org/10.1057/elmr.2008.73

7. Is there a role for data holders?

The preceding discussion could be interpreted as arguing that data holders have no role to play access to data once the data has been provided and acceptable scope agreed. This is not the intention. For the data holders, engaging with research projects can provide

- Statistical insights
- Methodological insights
- Education and upskilling of staff
- Opportunities for staff to share expertise
- New ways of thinking about policy issues
- Integration into research networks

For researchers, the engagement of data holders can provide

- Expert insights
- Policy insights
- Integration into policy networks
- Impact
- Partners for funding bids

Hence, the engagement of data holders with research provides benefits for both parties. The argument of this paper is that this should be a voluntary arrangements driven by shared interest, rather than a compulsory arrangement requiring shared responsibility.

Recommendations for UKRI

A2.9 UKRI should explicitly support the development of effective research networks including data holders and researchers, either directly or indirectly through data services.