Future Data Services
Data Access Theme Paper 1
Forms and process

Author: Felix Ritchie, Senior Strategic Fellow to FDS.

March 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.

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 1 *Forms and Process*

Contents

Summary of recommendations

Recommendations for data services

A1.1 The design of data access forms should follow principles of good survey design as the starting point; in other words, every question should satisfy the requirements (a) it is needed (b) it is not easily available elsewhere (c) the question can be understood and answered by the respondent (d) the answer can be understood and used by the data service.

A1.2 Data access form designers should consider the skills and motivation of those assessing the forms, to ensure that skills are sufficient and that motivation does not create perverse incentives

A1.3 All fields on the access form should be challenged and require justification before inclusion in the form; in particular, any claims that a piece of information is legally required need to be demonstrably evidenced by reference to legislation.

A1.4 Wherever possible, existing accreditation (Safe Researcher Accreditation, pre-existing ethical approval, DEA accreditation etc) should be used to avoid gathering the same information multiple times

A1.5 Collect data on problems with access forms, including time to completion and number of iterations

A1.6 Data access forms should be reviewed periodically, with the review led by someone independent of the design team/data service

Recommendations for ESRC

A1.7 Support research to understand what information is most useful in developing sensible, feasible and publishable research projects

A1.8 Publish and periodically review a good practice guide or templates, including international comparisons

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 issue of external approval is dealt with in the second of these papers. In this paper, we concentrate on the complexity of the process, which is very closely linked to the design of the application form.

The application form is central to process design as

(a) It embodies the thinking processes of the process designers
(b) It requires a degree of expertise (possibly none, possibly very high) from both applicants and assessors
(c) The ability for researchers to complete the form and assessors to assess it has a direct effect on the efficiency and resource needs of the process
(d) Form design affects users perceptions and hence their willingness to engage actively

2. Principles of data access

We begin by considering the fundamental principles of data access. A project proposal review determines whether a project is ethical, feasible, and cost-effective, but these determinations are strongly affected by the attitudes and processes of the review bodies.

2.1 Is it ethical?

Consider first whether something is ethical (not whether it is feasible or cost-effective). The relevant questions are

1. What is the benefit to the public of carrying out this project?
2. What is the risk to the public of carrying out this project (is the research activity or the final outcome likely to lead to disadvantage to specific groups)?
3. What is the risk to the public of not carrying out this project?
4. What are the privacy risks to the individual (natural or legal) who provided the data?

There are multiple philosophical approaches to considering how the benefit to the public compares to the risk to the individual ie (1) versus (4). However, in the case of the accredited Trusted Research Environments (TREs) in the UK for access to government and health data, this is simplified. We have ample evidence that the privacy risks to the individual of their data being used in these TREs are negligible. Formally, data within the TREs is treated as non-personal in GDPR terms because of the combination of procedural, technical and de-identification processes. Of course, there may be cases in the future where, for example, sensitive identified data are made available to researchers; but for now we can assume that for all practical purpose the TRE ensures that the privacy of the individual is not compromised. More generally of course, we know enough about data protection to make any user of the data effectively private for the purposes of a specific piece of research.

Once you remove (4) from the list, the ethical review is simplified. (1) and (2) should be on the application form. (3) is rarely asked, but it can be implied from as mostly the flipside of (1). From an ethical perspective, these should be very low bars: unless there is a downside identified in (2) you are only concerned with whether there is a meaningful upside from (1) and (3).

One could argue that reputational issues require there to be a sufficiently large 'public good' so that, for example, seemingly frivolous projects don't make it into *The Daily Mail* as an example of government waste, and energy has to be expended on dealing with that. But it's not that hard to identify academic research projects as contributing to some public good.

2.2 Is it feasible?

Feasibility concerns:

- Will the data be available to carry out the research (or a variant of it if the exact data turns out to be different)?
- Can the researcher carry out the project?

The first is a valid question for the review panel. If the user needs a dataset which is never going to be released, or requires linkages which are statistically useless, then the project is a non-starter. But this should be a very basic question about data availability.

The second question isn't relevant to project review unless it is very very clear that a researcher is manifestly unqualified and highly likely to produce results which would be damaging to the public interest. This is a theoretical problem but not a practical one. I'm not aware of any evidence that shows a researcher's competence, above a basic level, is a risk factor.

2.3 Is it cost-effective?

For public sector organisations, efficient provision of services is important. To set against the public good, the facility needs to consider

- The set up cost (acquiring data, creating environments, validating users)
- The ongoing costs (support)

The more idiosyncratic/bespoke/disorganised a service is the higher the costs. This is where the role of the project review panel moves into a grey area – should its decisions reflect the public good (and treat financial viability as a separate assessment process), or is the ability of the service to financially support the project a valid question?

2.4 Default open or default closed?

A default closed approach to data access means refusing access requests unless it can be demonstrated that they are of benefit and all the risks have been managed. It seeks to address the question "Can this be allowed?" and looks for problems.

A default-open approach to data access assumes that a submitted project is of worth and should go ahead unless it can be demonstrated that there is no safe, cost-effective solution. It addresses the question "How do we make this happen?" and looks for solutions.

The default-open approach works better because it puts the onus on finding a solution, turning this into a cost-benefit assessment. Under current UK law every *ethical* research question for which the data exists has a lawful route for research access. For example, if a project required access to 100% fully identified Census records, one can easily conceive ways in which the project could be managed securely. Those ways would also be very expensive, cumbersome and research-limiting, and so it is

hard to think of a research project that would have such overwhelming public benefits to make it worthwhile. But we can design a solution if that situation ever arises.

2.5 Exception, precedent, standardisation and accreditation

The most efficient organisations review by exception. The form, and guidance to scrutineers, should be such that most forms can be easily assessed, quickly, with relatively little knowledge, and recommendations made. This is necessarily crude, so the scrutineer (or the researcher) needs to be able to say "this needs a closer/longer/more expert review".

Ideally, almost all applications can be signed off automatically based on past precedent, speeding up the process for those and allowing limited resources to be concentrated where it matters. This is how principles-based output SDC works, and it is widely implemented in the UK social science TREs because it is both efficient and secure.

Operation by exception and precedent is made is easy if more of the process can be standardised and/or accredited. The less scope there is for variation, the more scope for precedents. For example, the DEA Accredited Researcher process (a pass in the Safe Researcher Training course, plus additional evidence of being a genuine and statistically trained researcher) means that there is no further necessity to check the credentials of the researcher, unless in some exceptional case.

3. What makes a good application process?

There are three stages to a good application process: forms, scrutiny, and communication.

3.1 Forms

When designing a survey questionnaire, there are copious guidelines on how to acquire useful information safely, appropriately and without bias. In summary, every question needs to pass three tests:

1. Is this question necessary – do we need this information, and if so, could we get it in another way?
2. Is this question comprehensible to the reader, given the level of knowledge we can expect of them? If it needs explanatory notes, are the notes in appropriate language?
3. Is the respondent able to give an answer in a way that is useful to you and in language that is meaningful to them?

This is well-understood in the survey literature, and should be the basis for all other forms, including applications for data access.

3.2 Scrutiny

We assume that there are two stages to an application process: scrutiny leading to a recommendation, and then a decision on that recommendation.

Those scrutinising the forms need to have sufficient knowledge to be able to do this. A way to reduce the knowledge requirement for scrutineers is to have questions which can be marked objectively eg using tick boxes instead of asking for a verbal description. For example, the UWE ethics form used to ask researchers to describe how their data would be stored, a question which was often badly answered. We lobbied to get the question changed to "will your data be stored in X, Y, Z, or some other place, and if the latter, why?" where X,Y Z are approved storage solutions. Since the question was changed, no researchers have chosen 'other' and almost all choose the preferred

UWE solution. So moving to tick boxes (or similar) can simplify scrutiny, but it can also be used to encourage positive behaviours.

Scrutineers need to understand what are significant problems, and what are not. Where can judgement be exercised? The more judgment, the more expertise the scrutineers need, and the more scope there is for inconsistency. However, without an allowance for judgement, the form either becomes too rigid for the range of research work, or too loose to meaningfully scrutinise. The circle can be squared by reviewing by exception (as described above), providing both speed and confidence. This means that scrutineers are not being asked to make a judgement on the quality of exceptional arguments, but simply on whether those arguments need to be referred elsewhere. This reduces pressure on scrutineers caused by concerns that a 'wrong' decision may be made.

Finally, decision-making in general and reviewing by exception are both simplified by having a clear mechanism, identifying, setting and recording precedents, in such a way that they reduce the set of future non-precedent decisions.

3.3 Communication

Researchers' activity is dependent on the application process. Clear communication about expected delivery times helps all parties plan and identify where and why delays occur. This includes delays outside the control of the application process (external assessment), and the delayed due to researchers not responding to queries.

In a precedent+exception model, median and quartile response times are more useful metrics than means.

4. What standards can be aspired to?

When the VML (the SRS' predecessor) was run by a research team, the published expected approval time was 48 hours, which was met. In exceptional cases (for government funded projects) applications could be prioritised. Approval for access was not tied to personal accreditation, which would be processed independently.

This was of course based on almost all requests being for academic or government funded research projects using existing datasets from the catalogue. This simplified approval processes as almost every application was covered by precedent. Applications went through a preliminary check for completion of fields by the (non-research aware) VML administrator, and then to the Microdata Release Panel (MRP) secretariat. The secretariat carried out a more technical review, returning to the VML team if any queries arose.

The full MRP reviewed almost no VML applications after the initial predecent-setting ones in 2003 (creation of the MRP) and 2008 (Statistics and Registration Services Act). There may have been a VML-specific application form, but I think the standard MRP application form was used, with a recognition that 'VML project' sufficed for most of the MRP queries about datasets, data storage/access, and person accreditation. Periodically the MRP team met with the VML team to review the process, usually to improve the guidance given to researchers.

The MRP was of course kept busy with other applications. However, as the VML accounted for a large proportion of applications but required almost no input from the main panel, the MRP was able to concentrate on these applications. This is how the precedent+exception model is supposed to work.

In the current UK framework for health and social science TREs, it appears that the most efficient performer is OpenSafely, which aims to clear all applications in a week (because the review panel only meets once a week). This does depend upon approvals from the NHS being supplied in advance, but these are usually fairly quick and require an exchange of emails between the researcher and the appropriate NHS data guardian. OpenSafely also benefits from having essentially just one data set. However, it also uses the validation of the system and of people to minimise the set of questions asked on the access form. As a result the application form itself is one of the shortest.

5. Linking to the Five Safes

The Five Safes is the dominant data governance framework in the UK public sector for social science and public health. It is also increasingly used in international contexts, with the same reach in Australia, Canada and New Zealand, and a growing prevalence in the US, and Europe.

The ESRC Future Data Services project has identified that the wider application of this framework provides improved messaging across the range of data services provision, including access. Work by the Australian Data Archive to link each of the five safes to an accreditation standard shows that is it feasible to do this, even though at present this does not exist in Australia.

In the UK however, for all TREs there is a clear accreditation system:

Projects	No independent transferable standard
People	Safe Researcher Training
Settings	DEA accreditation; NHS Security Toolkit
Data	No independent accepted standard of 'risk' yet. Explorations under way by Statistics Canada and various academic research groups (but see FDS Briefing note 1 *What is a TRE for?)*
Outputs	Output checking course for TRE staff run by UWE Bristol

In respect of an application process, the links between the five safes and necessary information required for approval are clear – see the appendix for workflow. HDR UK is already standardising on a Data Access Agreement for users structured around the Five Safes.

FDS (Access theme) Paper 3 *Access flows and accreditation* explores these issues in more detail.

6. What does an application form need?

On April 20th 2023, the ESRC Future Data Service team ran a workshop to review applications processes for data across the whole portfolio. The workshop was attended by researchers and data service providers, the latter covering a wide range of ESRC investments. UKSA was invited but declined to attend. However, the presence of the UK Data Archive meant that DEA access procedures were considered. The workshop reviewed multiple aspects of the application process, looking to identify good and poor practice, blockers and examples of effective delivery.

As the final exercise, the attendees were asked to define the 'ideal' application form. The criteria in section 3.1 above were proposed to them as the design principles. Table groups then came up with their own designs; FDS staff, with strict instructions to challenge any assumptions, acted as facilitators.

As a starting point, the participants were invited to consider a form consisting of

- Researcher names
- Project title

and to build up from there. The practicalities of implementation were not to be considered, nor were current practices, or the attitudes of other parties.

6.1 An 'ideal' form

The resulting application form, after collating and discussing results from the table groups, was as follows.

Field	Notes
Fields unanimously agreed to be necessary	
Research IDs	It was pointed out that names required further checking. An ID number for accredited researchers (such as the ONS Accredited Researcher Number, or a researcher passport scheme) should already contain all the information needed to contact the person and about their fitness to do the research. Table groups identified this independently, following the principle of 'ask for the minimum info you need'.
Project title	
Lay summary	
List of datasets required	
Start and end dates	Only suggested by one group, but an oversight by others – agreed in the discussion
Fields with some support	
Ethical approval already gained?	Ideally obviating need for further scrutiny; funding by UKRI/govt/others could be seen as proxy ethical approval
Technical summary of likely outputs	To confirm that the researcher has thought about the project, and perhaps give some measure for checking if the project is running as planned/within scope
Access method	Whether TRE or other solution
User end-point security	If not using a TRE – application processes for downloadable data also considered
Software needed	To assess feasibility – not widely supported as appropriate to do in the application process

It was also suggested that the application form should be structured around the Five Safes to improve communication. Other fields suggested which did not get support were

- Demonstration of feasibility
- Researcher names
- Methodology
- Where outputs would be published

6.2 Commentary

The most important thing is that the fields proposed are all unambiguous, should be easily answered by any competent researcher, and easily checked by a scrutineer with relatively little training. They also support the 'don't rewrite' aim by lay summary, description of outputs, and ethical approval all being re-usable from funding applications.

It is noticeable that no group suggested 'public benefit' needs to be made explicit, despite these being given as one of the example 'must have' fields in the pre-exercise briefing. It may be that the groups forgot it, or may be that they decided that the review panel should be able to identify public benefit from the above. Lack of time meant this wasn't discussed further, but if the latter reason this suggest the group strongly took on the idea.

There was some discussion about what to do if data owner approval was needed (eg for a new dataset). It was felt that this should be taken outside of the application process, perhaps checked in advance as for OpenSafely.

7. Case studies

In this section we showcase an example of good practice that does largely implement the principles-led approach above.

7.1 OpenSafely

OpenSAFELY (www.opensafely.org) provides access to electronic health records. It is a remote job server (users submit code and get results back, but cannot explore, i.e. visualise the patient-level data directly). OpenSAFELY was set up during the pandemic and remote working is the only option.

Applications for access have three stages (see https://www.opensafely.org/governance/os-workflow.jpg). Essentially:

1. Applicants submit a short online application form (covering content such as, title, description, confirmation of ethical approval, team, coding experience); the template form is publicly visible: https://docs.google.com/document/d/1ujG3OI2Q8zJz1kLPq6d2zMrBGX-zztEXghlsxn-0amk/edit
2. For research, applicants must obtain an HRA favourable ethical opinion; for a service evaluation or audit project, as well as institution ethical approval, a senior sponsor (such as a band 9 NHS managers, or national or local clinical lead) confirms that the project is viable, ethical and worthwhile; this takes the form of an email exchange
3. OpenSAFELY review panel (i.e. NHS England, the data controller) meets weekly to review and approve

OpenSAFELY encourages applicants to have conversations with the team (specifically senior OpenSAFELY researchers) before submitting applications for review, for example to discuss the project purpose being appropriately linked to the current OpenSAFELY COVID-19 legal purpose, or to assess if OpenSAFELY has the requisite data needed for the project. As a result, almost all applications are approved within a week of submission. NHS approval can take longer but again this is measured in days or (rarely) weeks rather than months.

The OpenSAFELY application process is simplified by effectively having just a couple of datasets. However, in many other organisations this is also the case, and applications take an inordinate amount of time. The difference is that OS has applied the minimalist rules applied above. In addition, as much as possible is managed separately; in Five Safes terms:

Projects	NHS approval from experts familiar with the process and research, focusing on the specifics of ethics and viability
People	Verified health researchers (e.g. evidence of safe researcher accreditation); verification is separate from the project application unless this is the first application
Settings	Independently verified and accredited
Outputs	Standard process and certified checkers
Data	Catalogue data

It is worth noting that OpenSAFELY is run by academic (some being clinical) researchers, who are strongly focused on the delivery of research subject to a confidentiality constraint. This mimics the original setup of the ONS Virtual Microdata Laboratory in 2003, which similarly was able to deliver more and faster services at lower resource than any comparable NSI facility. In addition, the OpenSAFELY team directly employs technical experts, with significant experience in electronic health record data, who have written the TRE codebase from the gound up; this close collaboration between users, OpenSAFELY researchers and technical experts further enhances the service's responsiveness to user needs, queries that arise and solving issues.

7.2 Other examples

Both the Australian Federal Government 'DataPlace' and the new Research Data Scotland data access model are potential other examples of the default open, princinples based approach.

8. Recommendations

The recommendations from this analysis are as follows:

Recommendations for data services

A1.1 The design of data access forms should follow principles of good survey design as the starting point; in other words, every question should satisfy the requirements (a) it is needed (b) it is not easily available elsewhere (c) the question can be understood and answered by the respondent (d) the answer can be understood and used by the data service.

A1.2 Data access form designers should consider the skills and motivation of those assessing the forms, to ensure that skills are sufficient and that motivation does not create perverse incentives

A1.3 All fields on the access form should be challenged and require justification before inclusion in the form; in particular, any claims that a piece of information is legally required need to be demonstrably evidenced by reference to legislation.

A1.4 Wherever possible, existing accreditation (Safe Researcher Accreditation, pre-existing ethical approval, DEA accreditation etc) should be used to avoid gathering the same information multiple times

A1.5 Collect data on problems with access forms, including time to completion and number of iterations

A1.6 Data access forms should be reviewed periodically, with the review led by someone independent of the design team/data service

Recommendations for ESRC

A1.7 Support research to understand what information is most useful in developing sensible, feasible and publishable research projects

A1.8 Publish and periodically review a good practice guide or templates, including international comparisons

Appendix: application workflow mapped to the Five Safes