FDS Briefing note 2

What is 'minimised data' for scientific research?

Authors: Felix Ritchie, UWE DRAGoN and Paul Jackson, Research Data Scotland

"Curiosity is a delicate little plant which, aside from stimulation, stands mainly in need of freedom."

– Albert Einstein, Autobiographical Notes (1949)

The issue

When data are prepared for scientific research, and when researchers request subsets of those data for their analysis, how much data should be processed at each stage?

The received wisdom is 'the minimum necessary for a specific research question'. This is often characterised as 'need to know', and is based on an assumption that this is the sole lawful setting. Such an assumption is problematic for legal, practical, scientific, psychological, operational and security reasons. In this briefing note we explain why this is the case, and how this leads to the (emerging) understanding of good practice.

For simplicity we focus on a Trusted Research Environment (TRE) deciding how to manage its data resources and access procedures. We use this example because TREs, by design, are safe places to archive data in the public interest at a detailed level sufficient to be able to fulfil any reasonably foreseeable scientific research question (see Briefing Note 1, 'What is a TRE for?'). The question is: what the appropriate level detail is at the point of researcher access? The TRE **can** be an archive of data and provide access to some, or perhaps all, of it to researchers, but **should** it? The arguments here can be extended to creating datasets for download.

We also focus on the use of UK government non-health data derived from individual records as the legal framework is clear. The analysis can be extended to health data derived from individual patient records, for example, without significant variation.

Legal - archiving in the public interest and scientific research

A TRE combines the functions of being a data archive and a provider of scientific research access to the data it holds.

The GDPR sets out how processing for archiving purposes in the public interest and scientific research should be understood and carried out. In its recitals a positive framework, including for linking data from registries, is laid out. The purpose of the Regulation is to ensure:

> *"…researchers can obtain **new** knowledge of great value with regard to widespread medical conditions [and] obtain essential knowledge about the **long-term** correlation of a number of social conditions [and] provide **solid, high-quality** knowledge which can provide the basis for*

> *the **formulation and implementation of knowledge-based policy**, improve the quality of life for a number of people and improve the efficiency of social services."*[1]

Here we have some important understandings. The knowledge to be gained is intended to be *new* – it is essential therefore that a researcher can be curious, and can pursue their questions into previously unexplored areas of inquiry. At the start of this process the destination may be unknown, but scientific research should not be inhibited from finding it. Over-specifying the research question for which data will be precisely prepared and access given risks fettering scientific inquiry.

The insights may be observable only in the *long-term*, hence the need to retain data in the long term and the need to retain variable which may become relevant in the future.

The knowledge needs to be *high quality*, from which we may infer that the input data needs to be of high in utility and hence in detail. Further, the knowledge needs to be *solid* and suitable for *implementation of public policy* – hence it needs to be reliable, repeatable, reproducible, and citable.

For these reasons, scientific research is not subject to the principle of purpose limitation (Article 5(1)(b), and the data processed must be adequate for scientific research with these qualities to be fulfilled (Art 5(1)(c).

As well as being adequate, the data processed for scientific research must also be relevant and limited to what is necessary. Together these construct the principle of data minimisation: "adequate, relevant, but not excessive". But the GDPR, for the removal of doubt, elaborates further for scientific research. Article 89 provides that if minimising the data would inhibit scientific research purposes then other safeguards such as organisational measures may be used in order to achieve the principle of data minimisation. This is crucial, and justifies the retention of and access to detailed datasets in TREs which provide those organisational safeguarding measures.

The special features of scientific research purposes as described in the recitals make it clear why this further elaboration of Art 5(c) is required:

> *"Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure **respect for the principle of data minimisation.** Those measures may include pseudonymisation **provided that those purposes can be fulfilled in that manner.**…"*
>
> *(Art. 89 s1; emphasis added)*

Article 89 goes on to say that technical measures (which may include pseudonymisation or other such technical measures) may be used to achieve the principle of minimisation *provided the scientific research purposes can be fulfilled having applied them.* In other words, the primary concern is the fulfilment of scientific research, and if minimising the data alone to achieve the principle harms research then organisational measures (the TRE measures) may be used as well - or even instead - to achieve the minimisation principle.

[1] UK GDPR, recital 157

In other words, the preparation of data for scientific research is determined by the foreseeable research activities[2]. The test is whether the fulfilment of the research will be impaired. If it will be, then the data are inadequate. But the data must be adequate, so the principle of minimisation must be achieved by supplementary organisational measures.

The last part of Article 89 might be the most important. It requires that if a scientific research purpose can be achieved through processing where identification of data subjects is prohibited, then the purpose shall be fulfilled in that manner. Again, this is a requirement that requires an understanding of context. The prohibition of identification can be achieved by de-identification, separation from the operational function of the data owner, and placing the data within the organisational controls of a TRE. When this is implemented, the data are not personal data when processed in that context. No obligations to the GDPR data minimisation principle persist when the processing is not the processing of personal data.

The Digital Economy Act

As well as the Data Protection Act 2018, suitable safeguards are found in the Digital Economy Act 2017 (DEA). In particular, the DEA requires (the "first condition") that the data disclosed to a researcher should be de-identified and unlikely to identify individuals when accessed through an Accredited data access provider (i.e. an Accredited TRE). The DEA does not otherwise cause an inhibition of scientific research by requiring the removal of details that could be reasonably be foreseen as being relevant to the research inquiry. The provisions concerning research make no reference to minimality of data. It does require (in s70 of the 2017 Act) the Statistics Board (i.e. UK Statistics Authority) to publish a Code of Practice[3]. The Code has six conditions (paragraph 2.4) for disclosing information; data minimisation is not one of them. It also has seven principles, which likewise do not require the research dataset to be minimised beyond the requirement of the first condition.

The crucial element here, unique to scientific research and statistics, is that 'data minimisation' can be achieved at least in part by organisational measures, because it is accepted that if minimisation was 'technical' only (i.e. data destruction) then some scientific research becomes unfulfillable. Fewer technical and hence more and balancing organisational measures allows the science to become feasible – and flourish. The GDPR is clear: data destruction to the extent that a research purpose is rendered unfulfillable not an acceptable solution, because such research serves a public good and should not be inhibited.

Practical issues

What is 'the minimum necessary'? To take a recent example, one author wanted to study the characteristics of labour markets at the local authority (LA) level using ONS' Annual Survey of Hours and Earnings. ASHE has data from 1986 to 2022, to postcode level, and with a large amount of detail about the job and the employer. In terms of the data needed for the research:

- *What years do I want?* This can be specified with reasonable accuracy – the project used five-yearly intervals as we wanted to look at change; but it might have needed additional years eg to look at the year-on-year effect of Covid

[2] It could be argued that this could be an objective determination: if a project sets out its scope, then certain records and certain variables are objectively required, and others are objectively not required.
[3] https://www.gov.uk/government/publications/digital-economy-act-2017-part-5-codes-of-practice/research-code-of-practice-and-accreditation-criteria

- *What sample do I want?* The study is looking at employment prospects of young people – so does it just need young people, and if so, what is the cut-off age? This might need to be determined from the data. Moreover, ASHE is longitudinal so exploring the effect of local labour market conditions during young worker's formative years on later-life labour market outcomes is potentially a valid way to determine what are the key factors in understanding early-years factors.
- *What geographical coverage is needed?* The quantitative research was part of a larger project on employment opportunities in the South Wales valleys. Other areas are necessary as a control group, but which? Only areas that have the same characteristics? A nationally representative sample? Or ones which match on specific factors? The determination of the comparator group is an empirical choice, even if there are strong theoretical reasons to suspect ex ante that one comparator group is more useful than another.
- *What variables do I need?* In theory, all of the variables in ASHE can be justified; in practice, only subset will be chosen. While there may be theoretical reasons for choosing a subset, a competent researcher will explore different options and specifications.

In short, the data that is relevant for research needs to be determined empirically, both because of the uncertainty of research, and because of the need to provide statistical justification for choices made. All this is obvious to anyone who has carried out their own research, but may not appreciated by those with an idealised view of what research involves.

Psychological issues

Faced with the practical issues of determining the relevance of data, researchers making applications for data will generally specify everything that they think *might* be relevant. This is the case even when applications processes are fast and effective. For something like DEA applications (currently taking up to 20 weeks), the time risk to a researcher of needing a project variation strongly incentivises them to maximise the data request. It is unlikely that an application reviewer would be able to provide a convincing case for a smaller dataset in opposition to the researcher[4].

Hence, requiring a 'minimal dataset' is almost certain to guarantee it won't be. In one facility recently reviewed by UWE DRAGoN team, data was created on a one-paper-one-project basis. At no point in the history of the service had a researcher come back and requested more data. This implies one of the following:

a) The research was limited because of lack of data being supplied
b) The researcher was able to specify in advance exactly the variables and sample required
c) The researcher requested more information than turned out to be necessary

Case (b) is the least likely explanation.

Operational issues

A minimal dataset argument implies that each research project has its own, specific dataset. This is not what is intended by currently legislation. Indeed, the SA's Code of Practice states:

> *Principles 5, Proportionality. Data must be disclosed or made available in a way that ensures the burdens and costs of doing so are proportionate to the*

[4] NHS data requests are typically reviewed by an expert familiar with the topic and data. Most UK data services offer support with defining project applications, but not with a view to controlling the researcher's choice.

> *anticipated benefits of the proposed research, regardless of who accrues the burden and costs. […] Data-holding public authorities are required to provide data as efficiently as possible, and to ensure that any cost recovery charges are proportionate to work undertaken specifically for the purpose of releasing data for specified research projects.*

Preparing the ASHE data for re-use by researchers is a one-off cost, whereas creating researcher-specific subsets requires additional costs. This former is the core operating model of ADRUK – supporting researchers to create a research-ready dataset which can then be offered through the Secure Research Service as an integral unit.

Another issue is replication/reproduction/peer review. Suppose researcher A asserts that serious crime is more prevalent in ethnic group 1 but they didn't use 'guilty/not guilty plea' because they thought it was irrelevant; researcher B, seeking to replicate that finding, should be able to factor in the plea that was entered to see if that changes the assertion of researcher A. This should not necessitate a whole new dataset's construction, and a new argument about why 'plea' is necessary. If nothing else, the replication dataset needs to be exactly the same as the original dataset in all respects with the exception of the 'plea' variable; otherwise it can be argued that different results are due to a different dataset, not a different model.

Security issues

Multiple versions of the same dataset can create confidentiality risks. Suppose one project studies all students, whereas another only focuses on those who have identified as 'male' or 'female'. From differences in the study findings, inferences may be made about the characteristics of non-binary trans students which would breach confidentiality guidelines.

This is not easy to spot. All core UK data services ensure that their staff are trained to look for this in outputs. Requiring data managers to also be aware of this adds an additional risk factor.

There are also security issues in linking data. It is feasible to assess the risk posed by linking two static datasets together with known fields and samples. It is much harder to assess linkage risk from dynamic datasets.

Both risks are generally very small/negligible compared to the risks of making the data available in the first place (and the Five Safes framework provide compensating controls; see Briefing Note 1). However, advocates of data minimisation typically do not identify or acknowledge them.

The philosophy of access: default-open or default-closed?

Guardians of confidential research data can adopt one of two attitudes:

- Default-open: data is available unless it is not feasible/cost-effective to manage confidentiality risks; access decisions focus on "**how** can we release this data safely?"
- Default-closed: data is not available unless confidentiality risks has have been demonstrably addressed; the focus for the guardian is considering "**can** we release this data safely?"

This has been extensively discussed elsewhere[5]. It is important because, although the two statements are logically equivalent (assuming some objective standard of 'safely"), the attitude

[5] Ritchie, F. (2014). Access to sensitive data: Satisfying objectives rather than constraints. Journal of Official Statistics, 30(3), 533-545. https://doi.org/10.2478/JOS-2014-0033

substantially effects the amount of data released: default-open gets more data out for research; the default-closed perspective hampers research.

Data minimisation is a concept independent of these two approaches. However, it demonstrates usefully where responsibility lies for taking decisions

- Default-open: those wanting to reduce the amount of data available need to show that (a) there is no impairment of research **or** (b) the risks of allowing that data to be used have not been addressed
- Default-closed: those wanting any data need to show that (a) the data is needed for research, **and** (b) the risks of using the data have been addressed

(Emerging) good practice

The answer to the question 'how much data should be licensed for a research project is "the dataset". In practice, most data services already operate this way, and have done for years, as this is cost effective and simpler for all concerned in all aspects of data and access management[6].

A counterpoint is to consider data services which do construct project-specific data files (including two recently reviewed by the UWE DRAGoN team). These are more likely to require more resources, to have slower application and access procedures, and to need more documentation.

There are examples of effective project-specific data, but these are not generalisable. For example, OpenSafely operates in this way, but its research data files are drawn from a live administrative database so the concept of 'the' data does not exist.

The idea that researchers should only get exactly the data they need has been around for many years, but it has had more prominence recently because of the references in the GDPR. As a broad ethical position the idea seems obviously 'good', but this brief examination has shown that the concept has strictly limited value and considerable flaws, and attempting to implement a naïve interpretation can cause unwanted side effects.

Proponents of 'need to know' sometimes create a straw man: "if we don't restrict data, the alternative is to let researchers have free rein over the data". Those who say that are failing in their Art 89 duty, which is to prepare the data organisationally and technically so that when the *further* use of data by the researcher happens it is not personal data. But more importantly, it is nonsense; there are well-understood conventions for preparing datasets:

- Remove data which appears to have little or no research value (eg contact information about respondents, or identifiers)
- Identify whether some variables/data subject may have a particularly high sensitivity (which could be for confidentiality or other reasons)
- Consider whether 'extra-sensitive' data can be changed in some way to reduce sensitivity
- Define the release dataset

Data services will often prepare multiple versions of the same dataset with different access restrictions (eg access to a particularly sensitive set of variable may require more justification);

[6] Some organisations (eg Statistics New Zealand) do provide researchers with random subsets of the data for projects; this is done as a confidentiality protection measure, not a data minimisation strategy.

Statistics Canada has been pioneering a dataset-user-facility mapping for its research datasets. This allows some fine-tuning of datasets for different purposes, but it is not done at the project level.

So a standardised dataset is not a free-for-all. It has been assessed to see what is **reasonably** needed for research **generally.** It is very efficient, and by allowing for a small and finite set of variations it can handle complex ethical requirements at least as well as fine-tuning to particular projects.