Research Data Centres:
The role of brokers for negotiating access to data

Richard Welpton, Cancer Research UK

richard.welpton@cancer.org.uk

Abstract

Detailed and confidential data collected by government agencies have traditionally been difficult to access by analysts and researchers, despite the advantages to society from doing so.

Access has, until recently, been constrained by technological and legal barriers; however, developments have reduced the impact of these.

Despite these developments, accessing detailed, and confidential data, for analysis and research, remains difficult. However, Research Data Centres (RDCs) provide a solution to reassure and efficiently distribute the risk of providing access to these data sources.

This paper explains how RDCs can act as a 'broker' for enabling access to confidential data sources.

Keywords **–** access to confidential data; RDCs

1 Introduction

Throughout the UK, researchers are increasingly requesting access to confidential data sources which hitherto, have been unavailable to them, for a variety of reasons. Often, these sources of data are collected, managed and held by government agencies. Protecting the confidentiality of these data, and the individuals in society who have contributed these data to the agencies, is essential to ensure that public trust in authorities that collect these data is upheld, in addition to complying with legal obligations.

The value of undertaking research using these confidential sources of data cannot be underestimated. Often, the additional level of detail (which elevates the confidential nature of these data) enables much more research to be undertaken than is the case when more 'anonymised' data are used. Examples include evaluation of government interventions and spatial analyses. Or, where data from different sources have been combined together, for example, linking of data from administrative sources that increase the number of characteristics about an individual or organisation, which enables more concise and robust research to be undertaken. Sometimes, the data subjects could be considered to be 'vulnerable': individuals at risk, or children, for example.

Typically, such data are 'de-identified', or 'pseudonymised', as the data do not contain directly identifiable information about the data subjects such as names and addresses; but with some effort, there is a risk that the identity of the data subjects could be revealed (because of the precise detail of the data which is useful for analysis and research).

The inability to analyse such data restricts understanding about our society; and government programmes may be executed without sufficient evidence about what works well and what does not.

Fortunately, the last ten years has witnessed a rapid growth in facilities (commonly described as 'Research Data Centres', RDCs) that enable access to confidential data (including detailed survey, linked and non-linked administrative data, biomedical data etc.) for analytical and research purposes.

This paper discusses the skills of staff and their contribution in running these facilities, with respect to acting as brokers: both in terms of managing access on behalf of data suppliers, and negotiating for access to data. As this paper will demonstrate, RDCs as brokers produce efficient outcomes for data access: a more equitable distribution of access to data; and better management and fair distribution of risk, which data owners can accept when making decisions about providing access to their data.

2 RDCs and establishing data access

As stated earlier, a number of RDCs now exist which provide access to confidential sources of data. These include the ONS Virtual Microdata Laboratory, the UK Data Service Secure Lab, and the facilities created by the Administrative Data Research Network. In the charity sector, both the Health Foundation and Cancer Research UK have established their own RDCs to manage confidential health records pertaining to patients.

A detailed description of how an RDC works is provided by Schiller and Welpton (2013). For brevity, it should be explained that RDCs operate on two principles:

1. Data always reside in the secure data environment of the RDC; the data never leave. Instead, statistical results are derived from the confidential data, and are released, subject to a check to ensure no individual could be identified, nor confidential information revealed (a process known as statistical disclosure control)
2. Access to data is distributed; rather than, data being distributed. Many RDCs now offer 'secure remote access' to data. The first principle above is met; but analysts and researchers can access the data at one or more destinations, without the location of the data changing

Established RDCs have demonstrable operating procedures, data governance protocols and accreditation to assure data owners, and the public, that they are capable of providing safe and secure access to confidential data. As a result, more confidential data sources, including detailed survey and administrative data, collected in confidence by data owners, are accessed through RDCs than ever before. Only a few years ago, it seemed impossible to envisage, for example, access to detailed linked education and earnings data. As a result, we have a better idea of the returns to education from studying different courses at different education institutions. Accessing more detailed health records has enabled us to achieve improved rates of, for example, bowel cancer screening uptake.

A well-established route exists for accessing detailed, confidential sources. Data owners supply their data to RDCs, who in turn, make access available to analysts and researchers. Processes are in place for managing the data, staff vet statistical results to prevent accidental re-identification of data subjects, before releasing to the analysts and researchers. The risk of providing data access (i.e. that data confidentiality is compromised) is distributed efficiently between the data owner, the RDC, and the analyst (the data owner no longer bears all the risk of providing access to their data).

As a result, relationships between data owners and RDCs have flourished; as successful secure access to confidential data sources have been demonstrated, and valuable analysis and research published as a result of this access, more data owners now allow their data to be accessed through an RDC.

3 Foundations of access

Prior to the development of RDCs, access to confidential data was less well established, and relied on good relationships between data owners and individual analysts and researchers.

Often, data would be supplied to an analyst or researcher directly from a data owner. Or else, an agreement would be made that the analyst or researcher would access the data on the premises of the data owner.

These 'one-to-one' arrangements continue to exist, and do so when both parties share a common objective, i.e. analysis which will benefit them both.

However, there are downsides to such arrangements:

1. The risk is unevenly managed (typically the data owner takes all or most of the risk of providing access to confidential data)
2. As a consequence, there is no incentive for the data owner to provide access to any other analysts or researchers, and the benefits that could accrue to society for making such data securely available are not realised
3. The arrangement may be informal, or opaque; which can have reputational consequences for the data owner (particularly in the event of an adverse data incident)
4. The analyst or researcher has a monopoly on the data; the ability to reproduce results which is standard practice in scientific research, is limited
5. Issues about data retention and metadata are unlikely to be addressed (neither data owner or analyst or researcher has an incentive to do so; unlike RDCs who experience 'repeat custom')

Unlike the RDC model of data access, it's clear from the description above that we achieve a much less satisfactory outcome for access to confidential data, and fewer benefits of data collection are realised.

Despite these limitations, we should still remember that these one-to-one arrangements do achieve data access; and can be a way of gently easing a data owner into the framework of longer-term and more robust access. It can also demonstrate to the data owner, the value of enabling access to confidential data.

However, it is unlikely that these one-to-one arrangements will usefully lead to long-term data access decisions, because of the risk and the amount of work on behalf of the data owner to provide access. Having separate conversations with a number of analysts and researchers about providing bespoke access is likely to be time-consuming (an important factor when one considers that data access is not often a priority for a data owner).

This is an area where an RDC (or data service) can help. The next section considers the role of an RDC as providing a 'data brokerage' service, acting as an intermediary between data owners and analysts and researchers to achieve efficient confidential data access outcomes.

4 RDC as broker

4.1 How can an RDC facilitate access to confidential data?

The concept of an RDC is now well-established. The ONS Virtual Microdata Laboratory, since it's establishment in 2004, has demonstrated how to provide safe and secure access to confidential data. The UK Data Service Secure Lab and Administrative Data Research Network are examples of RDCs that provide access to confidential data from a range of data owners (mostly government departments, but also, other research institutes that collect data).

All have adopted the 5 Safes framework (see Desai et al, 2016) for managing data; this demonstrates that RDCs have knowledge and experience of managing who is permitted access to data (and how they are monitored once access is granted); what projects are untaken with the data; that the technology and IT infrastructure where data are stored and accessed is robust; and that staff can ensure that statistical outputs, when released from the secure environment, cannot be used to re-identify data subjects, nor reveal any confidential information.

In this respect, RDCs can provide much assurance to data owners about how their data will be managed and accessed. They are better able to do so, as this is their incentive for operating: they have sufficient detachment from using the data themselves (unlike analysts and researchers), such that they will prioritise managing data and access to data. This is of benefit to all concerned, and especially to ensure that the flow of data available for analysis and research can continue.

A lone analyst or researcher is much less likely to be in the position to provide these assurances to a data owner: this is not their primary focus (which is to analyse data and produce statistical results).

RDCs, through their resources and expertise, can also unload the burden of risk which stems from providing access to confidential data. This is through their management of access to data as depicted by the 5 Safes framework; and in doing so, can enable the data owner to provide more access to confidential data for the long-term. Again, an analyst or researcher will not be in this position.

In turn, the advantages for analysts and researchers of an RDC are considerable too. In effect, it is the RDC that takes responsibility for the IT systems, secure transfer of data from the data owner etc. This leaves the analyst or researcher with the time to undertake analysis (although they still have a role in protecting the confidentiality of the data, but it is now less onerous).

In addition, RDCs can make access to confidential data available to numerous analysts or researchers, so the research community itself, and therefore the public, will realise the benefits of access to data more evidently than if access is solely directly between a data owner and an analyst or researcher.

4.2 The skills required

RDCs have to work to gain the trust of data owners before access to confidential data can occur. Putting this task in the context of mistakes which have occurred with confidential data, a wary public and adverse publicity (for example, laptops left on trains etc.), it is imperative that RDC staff can demonstrate beyond doubt, their robust approach to managing access to confidential data.

A number of RDCs now have some kind of information governance accreditation (for example, the UK Data Service Secure Lab and the Health Foundation both have gained and maintained the ISO 27001 standard).

In addition, it is important that data governance can be communicated clearly and succinctly: to data owners who may be pressed for time and are trying to make the best decision; to the analysts and researchers who will access their data (so they know the role they have to play); and to the public whose data are collected, accessed and analysed within the RDC setting.

Other skills required by RDC staff include:

- the ability to maintain positive relationships with analysts and researchers who use their services (to enable monitoring and foresee and deal with emerging issues

before they become problems, see for example, Desai and Ritchie (2009)
- training analysts and researchers before they access the data (and run through their legal access to data, how to use the service safely and effectively, and how to avoid producing statistics that could reveal the identity of data subjects and/or reveal confidential information).
- understanding of data (including legal requirements for protecting confidentiality, and the characteristics of data that make them 'confidential')
- statistical know-how in order to assess statistical outputs for the risk that they could inadvertently reveal the identity of a data subject, and/or confidential information about them
- a background in research and analysis, which enables staff to empathise with analysts and researchers, and ensure that optimal solutions for facilitating analysis and research with data can be achieved

These skills enable RDCs to provide an effective service, both for analysts and researchers, and the data owners who agree to supply data to the RDCs. Without qualified staff, the ability of an RDC to act as an intermediary and broker for data access will be limited.

This is aside to the data negotiation skills which typically are left to senior RDC staff: satisfying data owners and analysts and researchers that the RDC as the intermediate solution for data access is the most efficient.

Of course, much of this negotiation will involve the RDC demonstrating its potential to provide safe access to confidential data: this requires more than just the senior RDC staff and their negotiating skills; demonstration relies on all RDC staff demonstrating their competence.

4.3 Other advantages

Aside from distributing the risk of data access, RDCs perform other useful functions that directly relieve data owners of the burden of providing access to confidential data. These include:

- screening of applications to access to data, to ensure that potential analysts and researchers provide sufficient, succinct information justifying their access request
- managing the release of statistical outputs produced from data in an efficient manner
- production of metadata (information about the data sources, that can assist analysis and research with data, and provide information for applicants of the data)
- with sufficient knowledge transfer government departments can shift the bulk of on-going research data support to RDCs
- bringing together analysts and researchers, and data owners, together to discuss findings from the results derived from the data

These are all useful activities that many data owners would think of as time consuming, and have no provision for, but which are generally undertaken by RDCs.

5 Data broking approaches

As previously mentioned, often a productive outcome is achieved when a data owner and analyst or researcher first meet and share a common goal that requires using confidential data. Access can be provided, and it's a good first test to see if the data can be used for the purpose in mind.

However, this paper argues that such an approach will result in limited outcomes: access to data by one analyst or researcher, and with the data owner liable for most of the risk of providing access. Of course, it should be kept in mind that it is the analyst or researcher who is most likely to instigate the idea of using data, and can explain statistical concepts and what can be achieved with the data, and the relevance of the results to the data owner.

For these reasons, a sensible approach would be for discussions to begin between a data owner and analyst or researcher; with an RDC becoming involved when discussions about practicalities of accessing data, training, releasing statistical outputs etc., immediately and in the long-term, become necessary.

Or for example, the analyst or researcher could approach a data owner accompanied by RDC staff. This two-pronged approach brings together the expertise of the analyst or researcher, with the expertise for managing the data efficiently.

In any case, this paper argues that efficient access to confidential data is achieved when data owners, analysts or researchers, and RDCs work together in partnership.

6 Summary

This paper has considered the role of Research Data Centres (RDCs) with respect to providing a broker service between data owners (who may be risk-averse about providing access to confidential data), and analysts or researchers (who wish to access such data).

RDCs can play a number of roles in facilitating access to confidential data, providing that RDC staff have sufficient

skills to demonstrate that they are a robust service for acquiring, managing and providing access to such data.

The main role that an RDC can play is to efficiently distribute the risk of providing access to confidential data; ensuring that data owners do not take all the risk. With appropriate safeguards and assurances in place, RDCs can better facilitate access to confidential data, as they have the expertise, knowledge of data, experience of managing analysts and researchers effectively, and understand how to assess statistical outputs that are derived from the data, and released.

Not only does this provide reassurance to a data owner, but it reduces the workload for them considerably. Potentially, this enables more confidential data to be made available.

For analysts and researchers, the use of RDCs are clear too: they can receive more support when accessing the data, and the burden of supplying their own secure data infrastructure is lifted from them, leaving their time to be spent analysing the data.

For these reasons, RDCs can be thought of as playing a vital role in ensuring efficient access to confidential data occurs via their data brokering role.

Acknowledgements

The author acknowledges advice and feedback from Tanvi Desai, Director, Administrative Data Service; Felix Ritchie, University of the West of England. The author would like to thank the Administrative Data Research Network for sponsoring this submission.

References

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems.* Reading, Mass.: Addison-Wesley.

Schiller, D. and Welpton, R. 2013. *Distributing Access to Data, not Data*: IASSIST Quarterly, vol 38, no. 3.

Desai, T., and Ritchie, F. 2009. *Effective researcher management*: UNECE Work session on Statistical Data Confidentiality 2009, Bilbao, Spain, 2-4 December 2009.

Desai, T., Ritchie, F., and Welpton, R. 2016. *Five Safes: Designing data access for research*: Working Paper, University of West of England.