UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

26-28 September 2023, Wiesbaden

Smoothing the way for secure data access using synthetic data

Authors: Richard Welpton and Emily Oliver, Economic and Social Research Council (ESRC), UK

Richard.Welpton@esrc.ukri.org Emily.Oliver@esrc.ukri.org

Abstract

In the UK, sensitive and potentially disclosive data (including survey and government-owned administrative data) are kept securely and safely in de-identified form and are only accessible to accredited researchers through Secure Data Environments (SDEs). Using this data for research has enormous potential, although access can be constrained by the need for researchers to understand enough about these complex datasets for them to submit a viable project proposal, tensioned against the resource required for data owners to assess every application to use it, and data guardians to answer questions from researchers about the data. For the researcher, they need to be very invested to engage: they can't see the data in advance of applying for it; can't test it to see if it will answer their research question; it can take a long time to get hold of; and when they do, it might not contain what they need. It's also burdensome for the SDE as the researcher needs to spend a lot of time in the SDE exploring and preparing data ready for analysis. The resource costs to both researcher and SDE can be considerable.

Low-fidelity synthetic data can be an effective tool to improve the researcher journey because it can lower the barriers to understanding the data before giving researchers access to the real data. As well as accessing it for training purposes, researchers can use it for exploratory analysis to determine if the real data includes the variables they need. In turn, this can help support researchers to improve the quality of their applications for funding and data access; and to develop and test their code while they are waiting for access to the real data. Researchers can continue to develop their code outside of the SDE, therefore minimising the time and resources spent inside the environment. In the UK, only a small number of data services provide access to synthetic data, despite the development of numerous methods for creating synthetic data in the last decade or so.

The Economic and Social Research Council (ESRC, the UK funding council for social and economic research in the UK) has invested in a programme of work to support the creation and routine operationalisation to supply low-fidelity synthetic data to support data access for research and improve the efficiency of SDEs. This has been done largely through ESRC's Administrative Data Research UK (ADR UK) programme. They have:

- Conducted an in-depth study of the concerns and myths held by government data owners surrounding synthetic data production and use;
- Funded the creation of a Python Notebook tool to create synthetic data easily, at low cost and minimal risk which has been tested and approved by government departments;
- Formed a position statement across its UK partnership setting the vision for synthetic data within its wider remit and mission;
- Embarked on a significant project to explore the utility and use cases of different approaches to synthetic data creation and to evaluate the efficacy of different models to provide recommendations for how synthetic data production can be achieved at scale whilst still acceptable to data owners;

- Developed a public dialogue on the acceptability of synthetic data, and public understanding of it and its uses to increase trust and confidence in its development for research for public good.

This session will describe the secure data landscape within which synthetic data sits in the UK and explain the approach taken by ESRC and ADR UK to utilise it as a catalyst for better quality applications for funding and data access, and a smoother researcher journey. We will demonstrate the effectiveness of provisioning access to low-fidelity data by describing how it makes the researcher journey for accessing data and use of data in a SDE more productive, while simultaneously reducing the burden for data custodians and maintaining confidentiality.

1 Introduction

Considerable progress has been achieved to improve access to sensitive data for research, particularly in the UK. For example, the Office for National Statistics (ONS) launched the Virtual Microdata Laboratory (VML) in the mid-2000s (later to become the SRS – the Secure Research Service). In 2011, the UK Data Archive established the Secure Data Service (now UK Data Service Secure Lab). ESRC's ADR UK programme, a partnership between government and academic groups across all four UK nations, creates linked datasets from administrative sources, making these available to researchers through four Trusted Research Environments (TREs): SAIL databank (ADR Wales); NISRA (ADR Northern Ireland); eDRIS/Research Data Scotland (ADR Scotland); and ONS Secure Research Service (ADR England). These are all examples of Secure Data Environments (SDEs), also known as Trusted Research Environments (TREs).

These facilities have become common place across the health and social science research sectors because they offer a robust approach to accessing sensitive data. They reassure data owners that data they are responsible for, on behalf of the public, can be accessed safely (mitigating risk to individuals in the data) according to the principles of the Five Safes Framework[1]. SDEs are now considered the default option as far as access to sensitive data is concerned.

SDEs enable a range of data sources to be accessed securely. Consequently, researchers can better explain a range of health, social and economic phenomena. Examples of these data include:

- Business survey microdata available in the SRS and Secure Lab (these are sensitive due to the difficulty of anonymising the data and keeping enough utility in the data to undertake research)
- Detailed versions of social survey data also available in the SRS and Secure Lab, where the additional detail such as very low-level geographies or occupation codes not available in downloadable versions of the data offer new research insights.
- ADR UK has supported UK and devolved governments to make a range of administrative datasets available through their network of four SDEs. These data are de-identified, but not suitable for download because of their sensitivity, and offer utility for researchers.
- Health data such as cancer registration data and records from primary and secondary care services are accessible to researchers through organisations such as NHS England, and other SDEs.
- Linked health and administrative datasets are also now becoming available to researchers through ADR UK's network of SDEs.

The UK benefits from a legal climate that permits use of such data for research purposes; but culturally the use of the data described above continues to provide ethical and public perception challenges. Concerns about the misuse of data are understandably a constant feature of public debate in this area. This underlines the important role that SDEs have in maintaining the social licence to use these data for research in the public good. When managed through the Five Safes Framework, secure access to these data through an SDE provides assurance that such access leads to safe use in the public good.

Despite the SDE solution, it should be pointed that the cost of setting up and operating an SDE is high. Unlike distribution of data, secure access to data through an SDE requires:

- A technological solution (controlling access to researchers, data, projects; coupled with computational processing power)
- An auditable information governance and assurance framework
- Expert staff (technology, research, data management, statistical disclosure control etc.)

An SDE can only support as many data sources, researchers and research projects as its technology and staff capacity can allow. For example, between 2007 and 2010, the VML could support about 12 researchers accessing the facility simultaneously (the number of physical desks available at the offices where researchers could sit to visit the facility). When the Secure Data Service was launched in 2011, it could allow 40 researchers to remotely access the service at any one time; this was increased to 150 recently, following funding from ADR UK to expand and improve the service.

Other capacity constraints remain:

Inputs: procedures that researchers must navigate to access data in an SDE often require the researcher to explain in detail how they will use the data to address their research hypothesis. While metadata and documentation can help (when available), researchers often cannot describe accurately how they will use the data until they actually have access to the data. This creates uncertainty and can lengthen the application process.

Quality and completeness of information: occasionally, researchers who have spent considerable time gaining approvals for access to data discover that the data are not suitable for their research when they finally acquire access: a significant opportunity cost for them (and the data owner and SDE that support their access).

Outputs: In an SDE, researchers need to have their research outputs checked for potential disclosure before being released, a process known as statistical disclosure control. This is largely a manual process: SDE staff receive and process these requests. The ability to support researchers can be constrained simply by the number of staff available to service these requests.

Throughput: Much research involves exploring data and methods before a research question is answered. This iterative process relies on computing power to process data. In practice, little of this processing effort leads to a direct research output (for example, it may take several iterations to estimate a research model that yields statistical results that a researcher decides to publish). Yet depending on the technical architecture of the SDE, researchers may be competing for available compute resource, such as CPU/GPU memory.

One solution to address these constraints is to simply invest more money into SDEs, so more staff can be recruited, and more computational capacity can be sourced, etc. Despite such efforts in recent years, the demand to access these data sources continues to grow. SDEs are unlikely to be able to scale up to keep pace with this demand indefinitely.

This paper describes the potential of **synthetic data** to reduce these bottlenecks. We provide a vision whereby synthetic versions of sensitive data are routinely produced to:

- Enable researchers to assess data before making an application to access them; making sure they are the right data to support their research and help them accurately justify their use of the data when applying to access the data.
- Support the iterative process of research methodology and execution outside of the SDE and thereby reducing demand on SDE computational resources and demand for staff time to undertake statistical disclosure control (accepting the latter may be automated or partially automated in the future).

The next section outlines in more detail the challenges that researchers experience. We proceed by explaining ADR UK's efforts to pilot the generation of synthetic data, and then describe how these synthetic data can support researchers and enable SDEs to work more efficiently given their limited resources, resulting in improved outcomes for researchers and the policy world they support.

2 Challenges for researchers

Using sensitive data for research, such as administrative data, has huge potential, not least because there is so much of it. Administrative data, by its very nature, includes everyone. The datasets are enormous and complex, rich with potential for discovering insights about behaviours, trends, implications and consequences for individuals, communities and the policies and services they are dependent upon. By linking datasets and combining survey and administrative data, these insights that can be even deeper, and the things they can tell us can be transformational.

Access approvals can be slow to gain particularly for linked administrative datasets, because typically each data owner (that is, the government department, local authority or other public body) will want to approve requests. For the researcher this is dependent upon:

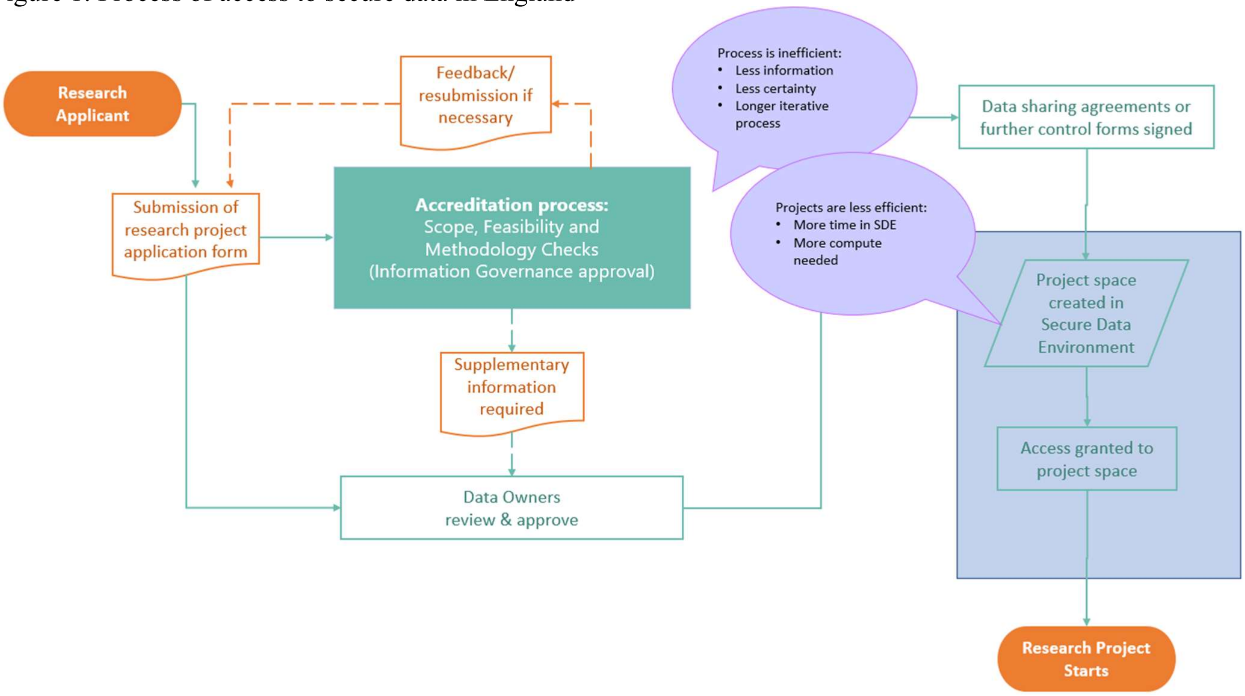
- Knowing what data they want to access – including the dataset, the variables, and even within the variable, the period of time they want to consider. Generally, a data owner will not want to give permission for a researcher to access any data they do not need to answer their specific question (the principle of only providing the minimum data necessary to address the research question). If the data has good, accessible documentation (metadata, user guide etc) this could be possible. Otherwise, they might need to rely on access to an expert who has used the data before and knows it well. The researcher needs to be specific and accurate in their request, but knowing enough about the data to do this before they make the request is not always possible.

- Getting a response from the data owner: this is dependent on the data owner having adequate resources in place to respond to data requests. The data owner needs staff who know and understand the data, who also have the time and remit to respond to these queries. If the data is deemed particularly useful by researchers and/or it does not have good and accessible documentation, the data owner might be inundated with requests for it. During times of political turmoil, such as during and post-elections, industrial action or national crises, processing data access application queries might be deprioritised.

Although a researcher can apply to be accredited to access secure data, it is generally only when the data owner has indicated approval can the researcher apply through more predictable channels: applying to the relevant SDE and getting confirmation from research approval panels.

Dr Paul Calcraft[2] has described the process of applying to access linked administrative data in the UK as trying to buy a second-hand car without being able to see it or test drive it first: Does it have all its parts, is anything missing, does it do what you think it will do, are there any quirks you should know about? In applying for data, one cannot see it in advance of applying for it; and one cannot test it to see if it will answer the research question. Accessing the data can be lengthy without certainty it will contain the information needed. Figure 1 sets out the process researchers need to follow to access secure data in England.

Figure 1: Process of access to secure data in England


Research Applicant
Feedback/ resubmission if necessary
Process is inefficient:
• Less information
• Less certainty
• Longer iterative process
Data sharing agreements or further control forms signed
Submission of research project application form
Accreditation process: Scope, Feasibility and Methodology Checks (Information Governance approval)
Projects are less efficient:
• More time in SDE
• More compute needed
Project space created in Secure Data Environment
Supplementary information required
Access granted to project space
Data Owners review & approve
Research Project Starts

3 Synthetic data as a solution

Bypassing much of the system for accessing secure data by instead accessing a version which is not real data and therefore does not need to be held securely, could be one solution for researchers. At the very least, using a synthetic version of the data to find out if you really do want to embark on a protracted process to access the real data, could be valuable. In this section we describe how this prospect should be considered.

3.1 Types of synthetic data and their potential

The utility of synthetic data for different applications is, of course, central to the question of its potential. High fidelity synthetic data which mimics the original data and preserves the statistical relationships between variables could reduce costs and complexities for the researcher, as it could also allow for analyses which are extremely close to those done on the real data. The use of such high fidelity synthetic data does come with a degree of risk for the data owner however, particularly if people misinterpreted findings from such data, or it was 'passed off' as real data.

Low fidelity synthetic data can, on the other hand, significantly reduce, if not remove, the risks for data owners, as analyses of the data would not generate meaningful results. It can also provide the researcher with easy access to a dataset which can be used to prepare code, test code, become familiar with the format of the data and learn how it can be used. It can also be used for training purposes, to raise awareness about the data.

In the UK, only a small number of data services provide access to any synthetic data, despite the development of numerous methods for creating it in the last decade or so. Making the production of low fidelity synthetic datasets more common could be beneficial to researchers and data managers alike. However, public perception of it is currently unclear and could be reputationally damaging if not addressed alongside other considerations.

For the purposes of this paper, we have described SDEs as 'remote access' solutions, in which the researcher can access and 'see' the data they have applied to access to undertake their research. Another approach is the ‘remote execution’ model, where a researcher develops statistical programming code using synthetic data, then submits their code to be run remotely on the data. Statistical outputs are then returned back to the researcher, subject to a statistical disclosure control check. Recent develops have included Application Programming Interfaces (APIs) to automate this process (such as DataShield, OpenSafely). Remote execution relies heavily on accurate synthetic data to ensure that the researchers can submit accurate statistical programming code; it may fail otherwise, to the frustration and delay of the research.

3.2 Developments

In 2020, ADR UK commissioned the Behavioural Insights Team (BIT) to undertake an investigation into the attitudes to and appetite for the provision of synthetic data by government departments. The intention was to understand the concerns and barriers to it with a view to being able to tackle these head on in a more informed way. It identified technical considerations, risk aversion and lack of knowledge, the use of advanced privacy-preserving technologies, and the need for better understanding of public attitudes to synthetic data alongside clearer communication as the key influencing factors. The results of the study are set out in the project report, Accelerating public policy research with synthetic data, and led to recommendations to:

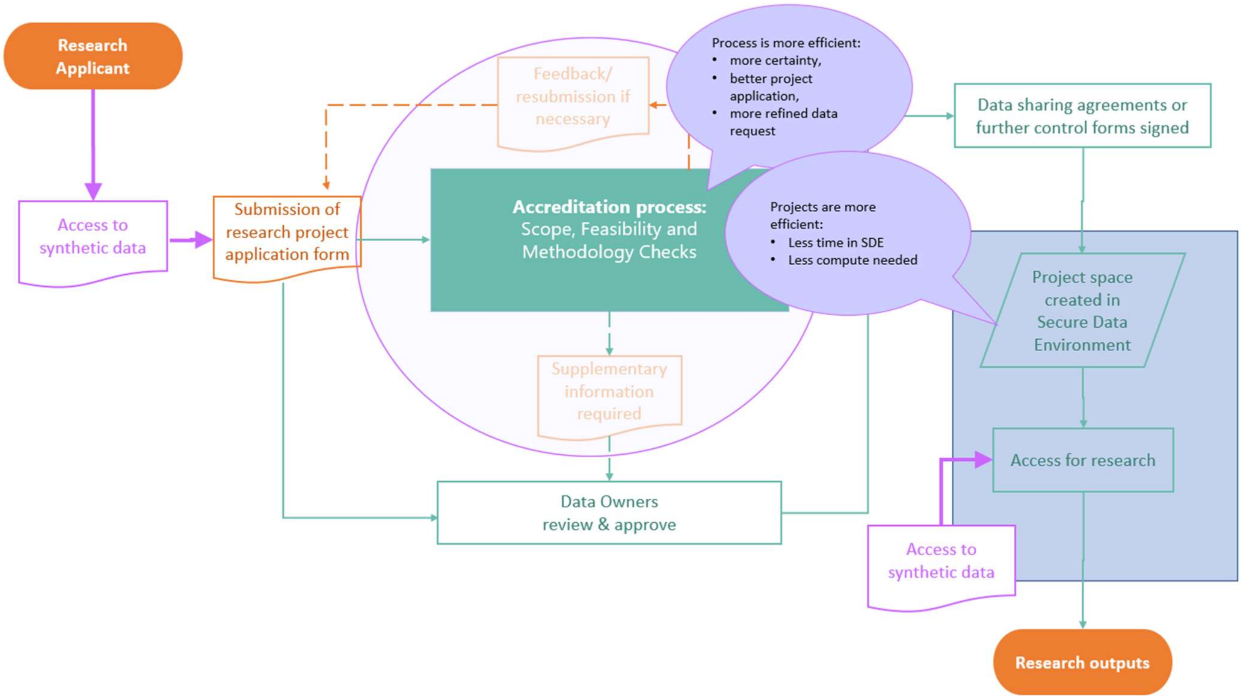
- Encourage the use and sharing of low-fidelity synthetic data to support rapid discovery of whether the dataset is appropriate for answering the research question; to develop and test code before full access is available; reducing delays in the process, including the amount of time needed to be spent in a secure environment;
- Expand the use of synthetic data for training so that researchers can be exposed to relevant idiosyncratic datasets earlier, thus improving their efficiency on live projects;
- Develop a cross-government repository of synthetic data for restricted access without a specific project proposal to allow for better design and more refined project proposals, and for this to be fed by a semi-automated pipeline to routinely generate low-fidelity synthetic data.

The study was followed up with the development of a synthetic data generation tool in the form of a prototype Python notebook which could be used by government analysts or researchers to generate low-fidelity synthetic datasets quickly and easily. It creates a version of the data that follows the structure and some of the patterns found in the real data. As such, it is plausible and represents the data as a whole. At the same time, because it does not preserve statistical relationships between columns, it reveals very little - if anything - about any individual in the dataset. The tool has now been extensively tested and is available for use. Users need Python (preferably Python 3), two common Python libraries (NumPy and pandas), and a software tool for viewing, editing, and running Python notebooks such as VSCode or Jupyter.

The BIT developers also produced a user guide which provides clear, step-by-step instructions, including how to ensure your system can run it. It guides the user through methods to run the cells in the notebook, explains how output files can be saved, and even tells you how to check that the notebook has worked. There is a useful section on troubleshooting as well as further information for more advanced users.

In an attempt to visualise the benefits of synthetic data for researchers using the process set out above in Figure 1, we have indicated on Figure 2 where the efficiencies could lie.

Figure 2: Proposed efficiencies on process when access to synthetic data is added:


Research Applicant
Access to synthetic data
Submission of research project application form
Feedback/ resubmission if necessary
Process is more efficient:
• more certainty,
• better project application,
• more refined data request
Accreditation process: Scope, Feasibility and Methodology Checks
Projects are more efficient:
• Less time in SDE
• Less compute needed
Data sharing agreements or further control forms signed
Project space created in Secure Data Environment
Supplementary information required
Access for research
Data Owners review & approve
Access to synthetic data
Research outputs

Of course, low-fidelity synthetic data is not a silver bullet. There will be instances where higher fidelity synthetic data is both more appropriate and more useful. In an ADR UK-led workshop at the International Population Data Linkage Network (IPDLN) conference in 2022, where different approaches to creating synthetic data were discussed, participants agreed that the value of different tools was entirely reliant on the end utility of the synthetic dataset[2]. Partners from ADR UK have taken their own approaches to developing synthetic data according to need and appetite in the devolved nations of Wales, Scotland and Northern Ireland and these have recently been published as an Interim Position Statement on Synthetic Data. It sets out ADR UK's vision for synthetic data and frames it in the wider context of its remit and mission. The statement is intentionally 'interim' because of the dynamic nature of this topic and our growing understanding of issues and opportunities associated with it.

3.3 Putting synthetic data into practice

While the case for the provision and use of synthetic data is powerful, data owners remain cautious, and we need to find effective ways of engaging the public in discussions about the creation of synthetic data. As such, we are a long way from seeing synthetic data operationalised to the point where trusted research environments can produce it routinely and facilitate access to it at scale. There is also a lack of evidence to support decisions among data owners and data services about how the governance around this might be best implemented. Data owners and services need real-world use case studies on costs and benefits to inform more systematic approaches to creation and sharing of synthetic data.

To inform future practice, ESRC and ADR UK are opening a joint research call to fund individuals and teams to explore how the potential of synthetic data can be harnessed at scale. Recipients of these grants will evaluate the current uptake, utility and governance of synthetic versions of datasets held in SDEs, including the benefits, costs and challenges to researchers, data owners and the SDEs themselves. They will also support a qualitative study of public understanding of and attitudes to synthetic data. The results of these funded projects will collaboratively inform a report and recommendations for how synthetic data production and provision can be achieved at scale and with the trust and support of stakeholders, including the public.

3 Discussion: Challenges and opportunities

The use of synthetic data provides an opportunity to reduce demand for SDE access, as analysis to complete projects within an SDE environment could be carried out more quickly. Our desire is that SDEs operate as efficiently as possible, and synthetic data, in our opinion, offers way to improve that efficiency, in the following ways:

It can enable researchers to make much more accurate data access applications. A benefit of this is that researchers will have more certainty that the data they are interested in accessing will support their research. Synthetic data should reduce the number of researchers who apply to access data, are set-up by the SDE to access data, but realise the data cannot support their research after all.

Researchers ought to be able to construct a significant amount of their statistical programming code outside of the SDE; and only use the SDE to refine and run the code on real data. This means they spend less time logged into the SDE and less time using compute resources for iterative coding.

If the use of synthetic data did create more opportunities to train and engage researchers in accessing sensitive data within an SDE environment, improve the quality of applications to access the data held, and also improve the efficiency of how SDEs operate, this may all drive up the use of this data for research in the public good. The process of producing useful synthetic data requires time, skills and customisation although much of the process can also be automated[3]. There are further challenges to address, including:

- Deciding which organisation is best placed to produce the synthetic data. The data owning organisation, or the organisation running the SDE?
- Should Digital Object Identifiers and other techniques be adopted to monitor version control and use of the synthetic data?
- What training and guidance should be made available to ensure that researchers do not inadvertently try to publish statistical findings that have been drawn from the synthetic version of the data, instead of the real data?
- How do we engage the public in discussions about the creation of synthetic data?

4 Conclusions

Synthetic data provides opportunities to smooth the researcher journey to access sensitive data via an SDE and reduce the burden on data owners and SDEs supporting researchers requesting such access. However, few use cases exist in the literature that evaluate the benefits and costs to stakeholders (researchers, data owners and SDEs), which is hindering scaled production and routine use of it. Evidence of public understanding and positive acceptance is not clear. Other barriers as described in this paper are not insurmountable and could, in the long run, reduce costs for stakeholders if automated systems were put in place. The benefits of the use of synthetic data are becoming clearer as more research is funded using secure data, the complexity of new, linked datasets increases, computational power increases and data science skills become better recognised for research across disciplines. For access to secure data to keep up with demand, synthetic data is a strong enabler and an important consideration for progress.

References:

1. Ritchie, F. 2008. Secure access to confidential microdata: four years of the Virtual Microdata Laboratory. *Economic and Labour Market Review, vol 2, No.5*
2. ADR UK Approaches to creating synthetic data: Workshop at IPDLN conference 2022.
3. Nowok, B., Raab, GM., and Dibben, C., 2017. 'Providing Bespoke Synthetic Data for the UK Longitudinal Studies and Other Sensitive Data with the *Synthpop* Package for R [1]'. Statistical Journal of the IAOS 33/3: 785 – 796. **DOI:** 10.3233/SJI-150153