Smoothing the ingest of data

Comparison across the 4 UK countries

Economic
and Social
Research Council

Version Control

Version	Date	Amendment	Author
1	05/06/2025	Report drafted by J Muir and amendments made by co-authors between 5th May and 5th June 2025.	Jen Bishop, eDRIS, Linus Chirchir, RDS, Rachel Coey, HBS, Amy Dunlop, NISRA, Dr Hywel Evans, SAIL Databank, David Grzybowski. SG, Alan Harbinson, HBS, Cristina Magder, UKDS, Nichola McCullough, NISRA, Richard McFerran, WG, Jen Muir, RDS, Alex Ramage, SG, Stella Telford, RDS, Richard Thomas, UK LLC Dr Tudor Vilcan, ONS SRS

Document Circulation / Readership

The intended circulation / readership for this document are as follows:

- ADR UK Operational Management Group
- ADR UK Conference attendees (poster abstract 126)

1 Table of Contents

2 Executive Summary

ADR UK's mission[1] is to bridge the gap between government and academia, to ensure good policy decisions from good quality evidence that addresses social and economic problems, leading to more effective public services and improving lives. This mission hinges on the preparation and provisioning of data for research.

Analysts across the ADR UK programme and beyond work to extract, process and curate public sector data and make it accessible to researchers in national trusted research environments (TREs). While each dataset, information asset owner, and organisation will be slightly different, the data processes can be made broadly replicable, with learning and best practice that can be shared across organisations.

ADR UK directed that an ADR UK Ingest Improvement Task and Finish Group be formed to explore the above. The initial purpose of the group was to summarise and evaluate data ingest processes across the four UK nations, identify lessons learnt, highlight areas of best practice, and inform potential improvements in the ingest processes. The scope and remit set out in the group's terms of reference was to focus on the data ingest processes from the point of extracting data from Government and/or public bodies into national TREs. To examine the processes for preparing the data for indexing, and any further processing required prior to ingestion into the TRE.

Membership comprises representatives from the Welsh government (WG), the Office for National Statistics' Secure Research Service (ONS SRS), UK Data Service (UKDS), UK Longitudinal Linkage Collaboration (UK LLC), Northern Ireland Statistics and Research Agency (NISRA), Health and Social Care (HSC) Northern Ireland's Honest Broker Service (HBS), the Secure Anonymised Information Linkage (SAIL) Databank, the Scottish government (SG), Electronic Data Research and Innovation Service (eDRIS) and Research Data Scotland (RDS).

The group have met online several times between March and May to share their processes, highlights and challenges. The group have a set remit but inevitably discussions touched on broader areas such as data acquisition as stages of the ingest process are interrelated. There have been three key areas that we collectively believe further work to address issues or develop common practices would highly benefit the ecosystem:

1. *Develop common coding standards and naming conventions*. We recommend code standardisation across the various teams working on data ingest and preparation, the development of a standard data ingest checklist, and a shared code repository on a platform such as GitHub or GitLab. This will support the scaling of work, allow for better understanding of processes and make it easier to share, reuse and maintain code.
2. *Improve infrastructure and processing environments*. The environments where data are curated often present issues and are not keeping up with demand, particularly when processing large datasets. Organisations have worked to adapt code and use different programming languages to attempt to address the challenges and improve processing times. Dedicating time and focus on the issue to develop common solutions, and allocating resource to implement any infrastructure changes will increase efficiency.

[1] What is our mission? - ADR UK

3. *Improve data owner and researcher relationships*: There are good relationships across the UK with data owners and researchers and these provide opportunities for learning that could be wider disseminated and trialled. Better relations can support both faster and wider access to data for preparing and making available to researchers, and improvements to data quality. Examples for fostering positive relations and engagement include:
    - In-person sessions with both data owners and researchers to discuss topics such as data issues.
    - Having a strong understanding of the governance and legal gateways surrounding data sharing and access for research.
    - Using case study examples to showcase the benefits for data sharing for research.
    - Creating synthetic data versions of all datasets for researchers to be able to assess which data and attributes they require for their study prior to application.
    - Beta testing new datasets by providing early access to new data for limited research studies. This can help identify issues that can be rectified before wider release of the data.

We recommend that ADR UK dedicate resources to address these key areas that promote the mission, and support and improve the ADR data provision community across the UK.

3 Comparison across 4 nations

The organisations responsible for preparing data for research across the UK operate differently. Some individual organisations cover the whole journey of data for research, from data ingest, provision, to outputs, whilst others carry out independent functions or tasks within the journey and securely pass the data on to another partner organisation.

There are also differences in approach to data provision, in that not all organisations ingest individual 'linkable' datasets but instead create 'themed datasets'[2] for researchers to apply for subsets of. These 'themed datasets' cannot be onward linked to anything else.

The organisations and functions carried out by each are as follows:

Organisation	Data Acquisition	Data Preparation	Data Provision	Researcher Access Service	Researcher Engagement
WG	✓	✓	✗	✗	✓
SAIL Databank	✓	✓	✓	✓	✓
ONS SRS	✓	✓	✓	✓	✓
UKDS	✓	✓	✓	✓	✓
UK LLC	✓	✓	✓	✓	✓
NISRA	✓	✓	✓	✓	✓
HSC HBS	✓	✓	✓	✓	✓
SG	✓	✓	✗	✗	✗
eDRIS	✓	✓	✓	✓	✗
RDS	✓	✓	✗	✓	✓

4 Areas of good practice

[2] Administrative Data Research Northern Ireland (ADR NI) themed datasets | Northern Ireland Statistics and Research Agency

The aim of UKRI's Digital Research Infrastructure Programme is to achieve a single national digital research infrastructure through interoperability, all built around the internationally-recognised FAIR data principles (Findable, Accessible, Interoperable, Reusable). The ESRC Research Data Policy outlines the roles and responsibilities of grant holders, and providers of research data, and expects compliance to the FAIR principles wherever possible.

There is a keen interest within the ADR UK Ingest Improvement Task and Finish Group in reproducibility with several member organisations having created open-source tools, modular code and Reproducible Analytical Pipelines (RAP). These methods evidence alignment to the FAIR principles. Further work to support wider adoption of reproducible practices and harmonisation like the use of common standards across organisations will ensure FAIR principles compliance and deliver efficiency-savings and scalability.

There are already significant efforts and developments to automate parts of the ingest process including data cleaning and disclosure checking. Some examples of good data preparation practices shared within the group are:

WG: Using modular code has improved accuracy, as sections of code can be reused across datasets. This gives consistent outputs from a verified code module. Maintaining the code is also easier, as duplication is reduced and updates are easily implemented.

Close relationships have been developed with SAIL to enable data inconsistencies to be flagged and investigated by WG, along with introducing SAIL's post ingest checks to WG code earlier in the process, so issues can be identified sooner, and the data quality report sent to SAIL as a benchmark.

Creating a conduit between researchers and data owners has created an informal feedback mechanism for issues that researchers may consider trivial but have been invaluable to data owners.

SAIL Databank: Upon data ingest, checks for: overall data volume and volume increase, duplicate rows, ID linkage rates, periods of missing data, any personally identifiable information. Additional dataset-specific checks are implemented for the most commonly used datasets which have regular refreshes. New data refreshes are compared with the previous version to identify issues. Work closely with many data owners to collate and share best practice.

ONS SRS: Semi-automated ingest checks are done via a RAP developed using the programming language, Python. This RAP can be applied to any dataset. It splits the dataset into different types of variables and creates different kinds of outputs for each. The checks carried out are for disclosure, not quality.

UKDS: Metacurate-ML and QAMyData tools (QAMyData open-source and available via GitHub, with Metacurate-ML to follow in 2026). QAMyData is a health check tool, helping curators and data owners understand common data quality issues and common disclosure control issues utilizing Regular Expressions and dictionaries to identify direct identifiers. Metacurate-ML is an automated metadata generation tool designed to produce FAIR-ready metadata from questionnaires. It uses advanced AI/ML techniques to extract and enrich questionnaire metadata. The tool also integrates with disclosure control processes by combining metadata with data content to support human-in-the-loop risk evaluation, using methods like sdcMicro. Additionally, at UKDS, in-house Python scripts are used to generate a range of dissemination formats, including SPSS, Stata, tab-delimited ASCII, and preservation-ready formats, as well as RTF data dictionaries. While these scripts are not open source, they are available upon request.

UK LLC: Metadata Explore tool and data pipelines for deidentified study data, with some modules generic for reuse with multiple datasets, and others specific to certain data. These pipelines clean and de-duplicate data, make sure table names and structures align with how they want metadata to be structured, and integrate variable value labelling if available. They also carry out an automated disclosure risk assessment, checking for things like dates, geographical identifiers, long and free text (free text is not allowed in the environment).

NISRA: Work on themed datasets and have a comprehensive related communications programme to promote and disseminate information and awareness of these datasets.

HBS: Each project specifies the minimum dataset required for their research. Data ingestion checks follow standard practices around checking against specification to ensure all variables are provided and complete, checking volumes are as expected and any specific instructions on the cohort creation have been followed. Robust disclosure checks are in place to ensure all PII is removed.

RDS: A RAP was developed using the programming language, R. It takes a modular code approach with different scripts to perform each preparation stage such as data ingest (from the raw data source to where the data will be processed), data preparation (including cleaning, validating and hashing), artifact generation (to produce dataset-relevant documentation), and checking (including manual elements).

eDRIS: Add geographical and deprivation variables via the patient's postcode which saves researchers having to continually derive these variables themselves. They have also launched a dose instruction project which is working to convert prescription text (e.g. 'take two, twice daily) to numerical values using machine learning. eDRIS also ingest data that does not hold Personal Identifiable Information but does hold location information specifically the Unique Property Reference Number, hashing the number as it is ingested to support the joining of people to place.

5 Key challenges

1. *Capability & skills*

Time and resource to develop new methods of working, along with the complexity of the data and tasks make it challenging to move to more automated, reproducible processes.

Team sizes and remit can vary dramatically across organisations, ranging from around 2 to 15, which impacts on capacity and responsibilities. Organisations can achieve more by collaborating to share methods and tools, making the most of the different organisation's expertise and advancements. This will allow others to learn and adopt efficiency saving practices within time or capacity restricted environment, avoiding duplication and reinventing the wheel.

2. *Maturity of Data Management within Organisations*

Synthetic data: Researchers will often apply for more data than they need because they are unable to discern what data they need from the metadata available. This leads to unnecessary provision of data, in contrary to data minimisation best practices. Making synthetic versions of datasets could help address this. There is work underway to develop an ADR UK strategy for synthetic data. As this develops, the generation of synthetic versions of the datasets we ingest will likely fall into our remit as ingest teams. This means our processes will need to be adapted. An example of progress in this area is Dementia Platforms UK where they are providing low, medium and high-fidelity synthetic data.

Data standardisation/harmonisation: There is a desire for the federation of data services and unified access across the UK. For example DARE UK are driving developments with their federated architecture blueprint, various driver projects, and the TREvolution programme. The NHS Research Secure Data Environment (SDE) Network is making advancements in federation. As these efforts grow, developing an approach to data standardisation or harmonisation will become more important and may fall within ingest team's remits. Examples of data harmonisation are mapping to Common Data Models (CDM) like the Observational Medical Outcomes Partnership (OMOP). It will be important to establish a way to handle differences across nations like within education. This is a vital area of work that will bring real gains to the research landscape, but it is important to acknowledge that it will likely have an impact on the ingest teams and require changes to process and/or additional workload.

3. *Technical issues*

There are technical issues around handling large datasets which result in adapting processes or offerings related to these. For example, within SRS, a range of tools and software are available to use for most datasets, but for large datasets this is restricted to SQL.

Processing speeds are an ongoing issue, and there are problems with infrastructure. Some examples are as follows:

- SRS state that the amount of data requested to ingest is a challenge, with admin data like LEO and eCHILD being around 1 terabyte each. Both require two disclosure checks of each variable alongside business as usual (BAU) survey updates at regular intervals. ONS SRS checks every dataset and each variable in the dataset for disclosure. This means checking that data does not contain personal information and there is no risk of secondary disclosure. Two checks are done for Quality Assurance (QA) purposes plus a smaller sign-off check which verifies the code outputs of the two checkers. This stretches

resources both in terms of people and time to carry out the checks, and computing due to the size of the larger datasets, but is necessary as safe data is cornerstone to the operation of the TRE.

- SRS are looking to machine learning use for numerical values checks, and a solution for non-coded string variables. They are working in an environment that is cut off from the internet which can make implementing new tools difficult, like the UKDS QAMyData tool. QAMyData has been made available as an opensource tool. It is written in a programming language called 'Rust' and requires a Rust Compiler which would be problematic to load onto, for example, a WG computer due to IT restrictions. This highlights a specific case where software harmonisation becomes essential to enable the sharing of best practice tools.
- RDS, HBS and others have worked to improve and update their processes to streamline code and extraction methods to tackle infrastructure issues and processing speeds.
- RDS have improved their code for working with large datasets which has increased efficiency, but further improvements would only be possible with higher computing capacity.
- HBS are working on process efficiencies for receiving files from the Regional Data Warehouse and are exploring Northern Ireland Health Analytics Platform (MS Azure) cloud functionality as a Secure Data Processing environment for preparation of research data. They have contributed to the development of a pseudonymisation[3] pipeline on this platform and are developing a standalone tool based on the open pseudonymiser for adding additional privacy protection when data is sent from external suppliers. The tool uses hashing and salting and enables the supplier, even if they're not technical, to pseudonymise data prior to submission.

4. *Significant relationships*

Building and fostering strong relationships with data owners and researchers will support faster access to data improve data quality and increase public trust.

Acquisition

Most if not all organisations struggle with encouraging some data owners to share data for research, a process of building and maintaining trust can take several months or even years to achieve.[4] This stalls the availability of valuable data that can be prepared for research in the public good and provide evidence to inform policy that can improve lives.

Data quality

[3] See the ICO's definition of psuedonymisation

[4] A useful tool used in data sharing discussions is the ADR UK document: *The legal framework for accessing data*

Finding the right balance between making changes to data that may improve its quality versus only providing informative documentation to explain nuances has been a challenge for organisations across the UK.

Where changes to data are made is crucial, for example by the data owner who provides the data at source or by the organisations processing data for research. There must also be a reliable mechanism for researchers to feed back issues they find in data they work with to improve the data quality. Ensuring a robust version control and communicating changes is extremely important.

Good practice in action can be seen:

- In Wales where they have different lines of communication to identify data quality issues that show the importance of building strong relationships to encourage informal feedback. Examples include:
    - ADR Wales work with SAIL to host quarterly researcher meetings attended by around 30 people[5], that invite data quality issues to be raised. These create the opportunity for informal conversations where researchers feel comfortable to raise data quality issues that can then be addressed.
    - Within the farm structure survey and the payments data, slight differences were seen in the county parish number variables. This caused issues when matching the data. Prefixes were added to identify which to disregard, and these were not identifiable after encryption. Quality issues were discovered because of internal relationships between the ADR Wales and AD|ARC teams which allowed the cause to be identified, the process understood, and the issue resolved.
    - SAIL facilitates written questions and answers between researchers and data owners to continually improve its data quality and associated metadata. A record of specific questions and answers are kept by SAIL to answer future requests without needing to consult the data owner again.
- In Ireland where they beta tested a new themed dataset by providing grants to around five projects to gain access for their study prior to the data being made widely available. The researchers involved fed back issues, and the team had time to resolve these prior to the dataset going live for all researchers to apply for.

[5] Of which there are around 15 researchers.

6 Recommendations

We discussed several challenges around capacity and skills, maturity of data management, technical issues, and relationships building. These each provide opportunities for further improvements and have been categorised into three key areas where we have made recommendations listed below. We have additionally provided detail on delivery approaches in Annex A.

1. *Common coding standards and naming conventions*

Implementing the recommendations below will support the disparate capacity of staff across the landscape and promote upskilling through shared learning:

a. **Recommendation**: *Develop code standardisation across the various teams working on data ingest and preparation to support the scaling of work, allow for better understanding of processes and make it easier to share, reuse and maintain code.*

b. **Recommendation:** *Develop a standard data ingestion checklist that covers the different stages of preparation and checks carried out on data for disclosure purposes.*

c. **Recommendation**: *Create an ADR UK shared code repository on a platform such as GitHub or GitLab.*

2. *Infrastructure and processing environment*

Recommendation: *Improve infrastructure and processing environments.*

3. *Data owner and researcher relationships*

Recommendation: *Improve data owner and researcher relationships.*

7 Summary and Next Steps

Between March and May 2025 this task and finish group has shared the ingest processes used within each organisation, showcased best practices and discussed challenges. This has allowed us to evaluate key areas to coalesce around and collectively recommend action. For the remainder of the group's lifecycle, from June until December 2025, the task and finish group will start to make progress on short-term elements of some recommendations listed below and described in detail in Annex A:

1. Common coding standards and naming conventions

    a. **Recommendation**: *Develop code standardisation across the various teams working on data ingest and preparation to support the scaling of work, allow for better understanding of processes and make it easier to share, reuse and maintain code.*

    b. **Recommendation:** *Develop a standard data ingestion checklist that covers the different stages of preparation and checks carried out on data for disclosure purposes.*

3. Data owner and researcher relationships

    Recommendation: *Improve data owner and researcher relationships.*

An additional allocation of resource will be required to address the longer-term pieces of work within the recommendations. We need to see all the recommendations that arise from the other ADR UK task and finish groups actioned in 2025 to understand any further alignments.

Annex A: Delivery approaches to recommendations

1. *Common coding standards and naming conventions*

Implementing the recommendations below will support the disparate capacity of staff across the landscape and promote upskilling through shared learning:

a. Recommendation: *Develop code standardisation across the various teams working on data ingest and preparation to support the scaling of work, allow for better understanding of processes and make it easier to share, reuse and maintain code.*

Summary: Encouraging and establishing consistent practices will promote processes that are more reproducible and easier to share, and this could link well with the new pan-UKRI policy as there is some additional work to be carried out around software and code[6].

Short-term: Within current capacity this group has until the end of 2025, we can explore clear folder structuring, modularised scripts and version control. We envisage this will be light-touch best practice sharing and discussions and individuals will be responsible for progressing improvements in their respective organisations.

Longer-term: Deeper alignment will be more time intensive and require an allocation of resource to pursue. This includes:

- Establishing consistent naming conventions.
- Incorporating tools that may be new for some including disclosure control tools or R Markdown and Quarto for reporting.
- Pursuing a UK-wide approach to data standardisation or harmonisation like mapping to OMOP CDM and handling standardisation across non-health data.

b. Recommendation: *Develop a standard data ingestion checklist that covers the different stages of preparation and checks carried out on data for disclosure purposes.*

Summary: There is commonality in the ingest processes across the teams. We suggest aligning these processes by developing a standard data ingestion checklist that incorporates shared best practice. This would mean the data ingested into our different TREs has undergone a standard preparation process and unified disclosure checks.

Short-term: We believe we can progress this recommendation within the remainder of the group's lifecycle. We can dedicate some regular meetings to share and correlate each of our ingestion checklists and agree alignment and content.

Longer-term: The standard data ingestion checklist developed and agreed by this group will be added to a shared code repository when available (see Recommendation 'c' below).

c. Recommendation: *Create an ADR UK shared code repository on a platform such as GitHub or GitLab.*

Summary: One shared platform will support harmonisation and coding standardisation, and teams to move to more reproducible, modular code. It will promote collaboration and help others

[6] Developing UKRI's research data policy

with challenges they've faced to make improvements. The code repository should have a definition of what it means by data ingestion and a good clear folder structure, split into the process themes with related documentation included. Any reference to such things as file paths specific to organisations should be removed before sharing.

Short-term: This recommendation will be achievable only with an allocation of resource to support the set up and management of the repository.

Longer-term: A shared code repository may have to initially remain private to the participating organisations, with the goal to either become open source or to grow an open-source element that will require additional resource for disclosure risk checking.

We believe researchers should be encouraged to adopt the coding standards agreed. Opening the repository platform to researchers as well as processors would be a useful move to further promote the FAIR principles and drive researchers to adopt the coding standards. Organisations can promote the standards as best practice and incorporate in data agreements with researchers and can document what's expected from researchers at the start. Organisations can trial/pilot this with some researchers once we have established a best practice solution. We would not expect researchers to go back and change their historic code.

2. *Infrastructure and processing environment*

Recommendation: *Improve infrastructure and processing environments*.

Summary: The environments where data are curated often present issues and are not keeping up with demand, particularly when processing large datasets. Organisations have worked to adapt code and use different programming languages to attempt to address the challenges and improve processing times. Dedicating time, resource and focus on this issue to develop common solutions will increase efficiency.

Short-term: This recommendation will be achievable only with an allocation of person and financial resource.

Longer-term: Resource can be used to support those looking at infrastructure solutions in different organisations to collaborate and establish if a common solution is appropriate. For larger scale improvements, resource will need allocated for infrastructure changes. This can include commissioning cloud platforms with for example Data Science Virtual Machines that provide additional processing power.

3. *Data owner and researcher relationships*

Recommendation: *Improve data owner and researcher relationships*.

Summary: There are good relationships across the UK with data owners and researchers and these provide opportunities for learning that could be wider disseminated and trialled. This will promote:

1. A mechanism for researchers to feed back issues they find in data they work with to improve the data quality. Feedback loops should include researchers, data owners and ingest teams for the most effective outcomes.

2. Effective communication that showcases the benefits and means for data sharing for research and supports the data acquisition function.

Short-term: There are elements that the representatives within this task and finish group can progress such as promoting this recommendation and related best practices to respective data acquisition teams. Specifically:

- Having a strong understanding of the governance and legal gateways surrounding data sharing and access for research. Representatives can share this ADR UK document used by some in data sharing discussions: *The legal framework for accessing data*.
- Using case study examples to showcase the benefits for data sharing for research. Promoting where research has had impact and/or saved lives. Impact case studies listed on the ADR UK website highlight the 'research impact' and can be used in discussions.

Some representatives will be able to trial researcher-related best practices like:

- In-person sessions with both data owners and researchers to discuss topics such as data issues.
- Instigating an effective feedback loop, using electronic communications that include all three parties: data owner, researcher and ingest team to alert all to the same issue or information.
- Systematically contact research offices, deliver presentations on researcher services and gather digital feedback.

Longer-term:

Synthetic data: We advocate for the creation of synthetic versions of all datasets for researchers to be able to assess which data they require for their study prior to application. This will foster positive relations. There is ongoing work in this space including a DARE UK funded UK Synthetic Data Working Group. An assessment needs made as to whether more resource is required to progress this element out with this task and finish group.

Case studies: Building a UK case study repertoire for different organisations to use to assist in data acquisition is beyond the scope of this group but would be beneficial. As referenced above, there are extensive 'impact case studies' on the ADR UK website and this may be adequate.

Beta testing: There has been success among partners like NISRA and SAIL Databank in providing early access to new data for limited research studies. This has helped identify issues that can be rectified before wider release of the data. It requires grant funding to pursue.

Education: Given the recently updated Information Commissioner's Office (ICO) guidance on Anonymisation, a unified awareness and education campaign targeting Data Owners would be useful in promoting data sharing for research and the enabling governance. The outcome of this effort would be more informed Data Owners and the value would come from reduced time in acquiring data for research purposes.