What is ‘research-ready’ data?

A report of the online roundtable event held on 11 February 2022

June 2022

Economic and Social Research Council

ADR UK

Overview

This report details a roundtable discussion hosted by ADR UK to explore what was understood by the term 'research-ready' when applied to administrative data. This discussion was informed by a recent systematic review of published literature on this topic (Mc Grath-Lone et al., 2022), and by the Data Utility Framework developed by HDR UK (Gordon et al., 2021).

Over 50 stakeholders attended the roundtable event including data owners, data processors, and data users. This report documents the:

- **key themes that emerged from discussions**
- **the range of different concerns and challenges raised by stakeholders**
- **suggestions for how these concerns and challenges might be addressed.**

The meeting was held on Friday 11 February 2022; a full list of attendees can be found in Appendix 1, and outstanding questions for future consideration are listed in Appendix 2.

1. Background

ADR UK (Administrative Data Research UK) is an ESRC investment which aims to transform the way researchers access the wealth of UK public sector administrative data for research that informs policy and improves lives. To achieve this, ADR UK funds projects that link administrative data from across different government departments to enable cross-cutting research for the public good, as well as funding data infrastructure and services to support secure access to these linked datasets.

Because administrative data is not originally collected for research purposes and often contains identifiable information, it must be prepared before secure access can be granted to researchers through, for example, a trusted research environment. This process of creating a version of an administrative dataset that is ready for research typically involves de-identification of the data (which is a requirement for accessing administrative data for research purposes under the Digital Economy Act 2017), as well as other cleaning and processing. Visit our website for more information about how ADR UK ensures data is used ethically and responsibly.

In the past, ADR UK has used the term 'research-ready data' in funding calls and in conversations with data owners as a catch-all term to communicate the processing of administrative data that is required before they can be made available for research. Through funding data linkage projects, it has become apparent that the concept of research-readiness has varying interpretations for different stakeholders. This point also emerged from a recent systematic review (Mc Grath-Lone et al., 2022), which found that the use of the term 'research-ready' by researchers varied: for some, it was used to describe well-defined data that was analysis-ready, or 'plug and play' (for example, query-able databases), while others used the term to describe datasets that were broad and less curated ('warts-and-all') and allowed the researcher the flexibility to prepare the data for different research purposes.

Research-readiness is a critical issue for administrative data because it often takes the researcher and data custodian a lot of time and effort to negotiate access to the data, which is wasted if the data are then not fit for purpose. As a proponent of 'research-ready data', we felt ADR UK needed a clearer understanding of how key stakeholders interpreted this term, with a view to developing a consensus or shared approach. Stakeholders include data owners, who typically perform the initial data cleaning and often analyse the data themselves; trusted third party data processors (like the Office for National Statistics), who might de-identify and link the data and facilitate access for research purposes; and researchers.

1.1. *Format of the event*

- Welcome and context setting – Emma Gordon, ADR UK
- Presentation: What makes administrative data 'research-ready'? - Louise Mc Grath-Lone, UCL
- Presentation: HDR UK's Data Utility Framework – Ben Gordon, HDR UK
- Breakout room discussions, steered by the following questions:
    - How 'good' (clean, curated etc.) does data need to be to be useful?
    - 'Transparency' of data: What are the barriers to making broad, messy data available for research?
    - What mechanisms are needed to iteratively improve datasets through research which uses them?
- Feedback from breakout rooms
- Round up and next steps

2. Key themes from the presentations and break-out room discussions

2.1. *Research-readiness has multiple characteristics, or 'dimensions'*

The systematic literature review by Mc Grath-Lone et al. (2022) identified five characteristics that define an administrative dataset's research-readiness: accessible; broad; curated; documented; and enhanced. HDR UK's Data Utility Framework also described datasets with very similar characteristics, or 'dimensions', including around access and curation. A key theme from the roundtable discussions was that **there is no single measure of what constitutes a 'research-ready' dataset and the measure of research-readiness of a dataset will differ depending on what it is being used for**. Because different users will want to use data for different purposes, they will have different thresholds for considering data 'research-ready'.

> "More curation is not always better"
>
> Roundtable participant

A dataset that is research-ready for one research purpose may not be suitable for another. One participant suggested that research-readiness of a dataset will also depend on what data already exists in that space. For example, because housing data related to private renting is very sparse, even comparatively poor-quality data within this theme is valuable. Participants also highlighted that a more curated dataset is not always more useful for everyone: some researchers prefer to access less curated data where they have more control over and input into the decisions involved in preparing the data for their particular analyses.

The perceived research-readiness of data will therefore also depend on the technical skills of the researcher (with more experienced researchers generally favouring broader, less curated datasets). Ben Gordon explained that because user requirements for datasets will differ, HDR UK has adopted the term "data utility" in their framework, with the aim to generate a common language to describe a dataset's potential for use for different purposes.

2.2. Research-ready datasets develop over time

From the discussions in this session, it was also apparent that 'research-ready' datasets are not one-off creations. Instead, datasets tend to develop over time along the identified 'dimensions' of research-readiness as they are used more frequently for research purposes, and users are given opportunities to contribute to the body of knowledge that accumulates around a dataset. HDR UK's Data Utility Framework uses: data documentation; technical quality; data coverage; and access and provision as its dimensions. The framework categorises datasets along each of these dimensions by defining varying levels of maturity, with datasets categorised as bronze, silver, gold, or platinum.

The development of research-readiness over time allows the burden of data curation to shift away from the data owner (or a trusted research environment) to be shared with research users who can feed back to data owners on usability aspects, such as the quality of coding within the data. This feedback loop was illustrated by one participant who described how social services data in Northern Ireland, which has only recently been made available for research purposes, is typically less clean than health data which is an established research resource. The historic use of health data for research purposes has enabled feedback between researchers and the data owner, including on what the data can or is being used for, and where the quality issues are. This research-user feedback in turn can help the data owner improve data quality at the collection, input, and data cleaning stages before making the data available for research purposes.

3. The concerns and challenges raised by stakeholders

3.1. Challenges for data custodians to make their data research-ready

- **Standardisation:** within government departments, often individual teams clean up the data for their own purposes, which makes standardisation, even within a data owner department, challenging. For example, one data custodian commented that there may be a distinction between an 'operational systems team' and the 'analytical data warehouse' that takes ingests from operational systems. Data for operational purposes is very different to data required for analytical purposes, so a lot of work is required before analysts get access to 'raw' data, which is far from raw, but also far from research-ready.
- **Resources:** preparing and cleaning data is resource-intensive and requires both technical and substantive expertise. This is typically non-trivial on both counts. For example, one partipant commented that the Department for Work and Pension's pensions credit benefit data is derived by extracting data from over 50 datasets. Summarising the data and taking away some of the 'noise' to prepare it for research requires extensive contextual understanding of how the data was collected to make sensible decisions on how to handle it. Sharing this data curation burden with researchers by making messier data available won't necessarily relieve this resource issue if the researchers then have lots of questions for the data owners about how the data was collected.

> "Who owns the burden of explaining the messiness of the data to researchers?"
>
> Roundtable participant

- **Lack of metadata for historic data:** there may not be information available about how data was collected or derived for historic data which inevitably will create questions for the researcher which cannot be answered easily.

3.2 Challenges for researchers to improve datasets to make them more research-ready

> "Industry is also grappling with many of these issues; are there insights we can draw from there?"
>
> Roundtable participant

- **Resources**: researchers are rarely rewarded for sharing resources that contribute to the curation of research-ready administrative sets, such as well-annotated code that they've used to clean or prepare a dataset for a particular purpose or descriptions of metadata. There is a need for a culture shift to incentivise these activities and highlight why they are important.
- **Responsibility**: one breakout room identified the need for this to be a community effort, where trusted research environments as data infrastructure providers could play a role in helping researchers to share their code with others.

3.3 *Challenges for trusted research environments to support the development of datasets by researchers*

- **Code sharing**: sharing code crops up frequently in discussions about how to make research on administrative data more efficient. Code that has already been developed can be saved and reused or developed. However, to do this, there is a need to develop the technical infrastructure to support code sharing within the secure environment. Additional resources may be required to check or quality assure user-produced code for cleaning and analysing data.

4. Stakeholder suggestions for how challenges related to research-ready data might be addressed

4.1. *Better documentation, and better communication of a dataset's utility for different purposes*

The roundtable discussion highlighted that 'documented' (a key characteristic underpinning the accessible, broad, and curated categories identified through the systematic review (Mc Grath-Lone, 2022)) was of critical importance to overcoming challenges related to developing research-ready data resources. Breakout room discussions also revealed that documentation is critical for researchers in the discovery phase of a research proposal so that they can scope their research questions effectively and judge the utility of the dataset relative to these questions before they go through the often-lengthy permissions process to access the data.

As well as better documentation, stakeholders identified the need to be able to communicate a dataset's utility for different research purposes. For example, categorisation of datasets along data utility dimensions in HDR UK's Data Utility Framework allows a researcher to quickly see whether the data is (un)suitable for their purposes. The Heath Data Research Innovation Gateway is also testing a Data Utility Wizard which allows a researcher to filter their searches for data by their different requirements (e.g., a researcher needing datasets that span a long period of time might use the Wizard to filter for datasets that include data going back at least 10 years).

> "Documentation isn't just about what's in the data, but also how it's collected and its [original] purpose"
>
> Roundtable participant

Finally, a number of participants identified synthetic data as a powerful tool to help a user quickly get to grips with a dataset's characteristics and utility for their research question with minimal disclosure risk and associated access requirements.

4.2. Routine access to different versions of datasets and any associated audit trails

Routine access to different versions of research-ready datasets would resolve the tension between some users wanting access to the broadest and least curated data, and others needing more focused and prepared data. However, access to different versions of datasets needs to be underpinned with access to documentation that describes any processing or preparation that has been undertaken, to ensure researchers have a clear understanding of how datasets have been derived (an audit trail).

This could include access to all the annotated code used to transform the data at each iteration or for different research purposes. This would enable complete transparency in the processes that take the messiest, raw dataset through to highly processed versions that are 'analysis-ready' for a particular research question and provide access to variably curated versions for researchers depending on their requirements. Because users often require the data cleaned in similar ways for different analyses, this would also greatly reduce duplication of effort.

> **"Data is not the truth, but a signal of the truth"**
>
> Roundtable participant

As one participant commented: "*Data is not the truth, but a signal of the truth*". That signal can be muddied by assumptions or variation in practices in the way the data is collected, or subsequently transformed. How data is handled through de-identification, cleaning and then linkage has often been a black-box. However, new computational methods and tools now allow the opening up of this black-box by enabling complete transparency through each computational handling. Consequently, **knowledge and transparency** (through coding of that knowledge) in how data has been generated and handled from collection right through to final analysis is critical.

4.3. Support and funding for intermediary bodies to curate and disseminate administrative data on behalf of data owners

The creation, curation and dissemination of research-ready datasets will require ongoing support, including funding. Intermediary bodies taking on these roles would ease the considerable resource burdens on data owners.

Conclusion

The level of interest in research-ready administrative data from stakeholders is considerable, indicating that there is a need to keep this conversation going and to offer more opportunities to engage with this evolving topic. It is clear from the discussion at this event that central to any opinion about how 'ready' data needs to be before it is useable for research is the importance of documenting the changes data goes through as it journeys between data owner and end user. This needs to occur both as a paper trail showing what has been done and how historic versions can be accessed, while allowing users to feedback on the quality and utility of these as they use the data to explore the insights it can provide. This finding gives us a useful foundation for future conversations and consultations, and points to a benchmark for development. If we can drive up the acceptance of and participation in enhanced audit and documentation of data, we can envision a culture change that supports more efficient research opportunities using administrative data.

This roundtable event positions ADR UK to consult further on this topic to refine and consider more deeply what we mean by research-ready data as data becomes available to, and is used by researchers, and through wider consultation. We intend to continue facilitating ongoing conversations with existing stakeholders and will also look to colleagues abroad to understand interpretations of what research-ready data exists in other countries and regions. The outstanding questions collected during our discussion (presented below in Appendix 2) will form a basis for ongoing work on this theme.

References

Gordon, B., Barrett, J., Fennessy, C., Cake, C., Milward, A., Irwin, C., Jones, M., & Sebire, N. (2021). Development of a data utility framework to support effective health data curation. *BMJ health & care informatics*, *28*(1), e100303. https://doi.org/10.1136/bmjhci-2020-100303

McGrath-Lone, L., AJay, M., Blackburn, R., Gordon E., Zylbersztejn, A., Wiljaars, L., Gilbert, R., What makes administrative data "research-ready"? A systematic review and thematic analysis of published literature. *Int. J. Popul. Data Sci.*; 4. 10.23889/ijpds.v7i1.17182022. https://doi.org/10.23889/ijpds.v7i1.1718

Authors

Balint Stewart, Research manager, ADR UK, balint.stewart@esrc.ukri.org

Emily Oliver, Head of Research and Capacity Building, ADR UK, emily.oliver@esrc.ukri.org

Louise McGrath-Lone, Research Fellow, UCL, l.mcgrath-lone@ucl.ac.uk

Visit the ADR UK website

@ADR_UK

Appendix 1: List of attendees, ordered alphabetically by organisation

Chair: Emma Gordon**, Director, ADR UK**

Stakeholder	***Representatives***
Administrative Data Research Centre Northern Ireland	Dermot O'Reilly, Director
Administrative Data Research UK	Emily Oliver, Head of Training & Capacity Building
Administrative Data Research UK	Balint Stewart, Research Manager
Administrative Data Research UK	Rosie French, Deputy Director
Administrative Data Research UK	Christine Boase, Strategic Lead for Campaigns and Communications
Administrative Data Research UK	Gregory Meredith, Senior Research Manager
Behavioural Insights Team	Bobby Stuijfzand, Research advisor
Consumer Data Research Centre	Maurizio Gibin, Technical Center Manager and Senior Data Scientist
Consumer Data Research Centre	Paul Longley, Director
Consumer Data Research Centre	Peter Baudains, Research Data Scientist
Department for Education	Gary Connell, Head of Data Ownership and Data Sharing
Department for Work and Pensions	Emma Slater, Central analysis and science directorate
Department for Work and Pensions	Graham Knox, Central analysis and science directorate
Department for Work and Pensions	Andrew Needham, Senior Analyst
Economic and Social Research Council	Richard Welpton, Head of Data Services Infrastructure
Health Data Research UK	Ben Gordon , Executive Director, Hubs and Data Improvement
Health Data Research UK	Jackie MacArthur, Research Project Manager at BHF Data Science Centre
Health Data Research UK	Susheel Varma, Chief Technology Officer and Director of Engineering
Health Data Research UK	Varsha Khodiyar, Data and Connectivity Project Manager
HM Revenue and Customs	Richard Millington
HM Revenue and Customs	Tracy Holland, Data Sharing Front Door Lead, CDIO Data Sharing and Acquisition
HM Revenue and Customs	Kevin Lindoe
HM Revenue and Customs	Angela Martindale, FP&P CS Security & Information Team
London School of Economics	Polly Vizard, Associate Professorial Research Fellow
Ministry of Justice	Kylie Hill Research and academic engagement lead, Data First
Office for National Statistics	Ore Odusanya , Project Manager
Office for National Statistics	Tom Carr, Head of Data, Secure Research Service
Office for National Statistics	Rachael Colquitt, Head of ADR Data Acquisition
Office for National Statistics	Bill South, Head of SRS Data & Governance
Scottish Government	Cecilia MacIntyre, Statistician

Scottish Government	Nicola Kerr, Data Sharing and Linkage, Digital Directorate
Scottish Government	Thomas Alexis, Statistician
Scottish Government	Ross Waddell, Assistant Statistician
Scottish Government	Katherine Falconer, Head of Information & Analysis, Registers of Scotland
Scottish Government	Scott Mcfarlane, Assistant Statistician
Swansea University, Population Data Science	Ashley Akbari, Senior Research Manager & Data Scientist
Swansea University, Population Data Science	Fatemeh Torabi Research Officer and Data Analyst
Swansea University, Population Data Science	Pete Arnold, Senior Lecturer
Swansea University, Population Data Science	Vesna Vuksanovic, Senior Lecturer
University College London	Louise McGrath-Lone, Research Fellow
University College London	Pia Hardelid, Senior Research Associate
University College London	Ruth Gilbert, Professor of Clinical Epidemiology
University of Bristol	Rosie Cornish, Research Fellow
University of Glasgow	Mark Livingston, Research Fellow
University of Hertfordshire	Tim McSweeney, Senior Lecturer
University of Manchester	Julia Kasmire, Research Fellow
University of Warwick	Arun Advani, Assistant Professor
University of the West of England	Felix Ritchie, Professor of Applied Economics
Urban Big Data Centre	Andrew McHugh, Senior Data Science Manager
Welsh Government	Anthony Whiffen, Senior Statistical Officer
Welsh Government	Tanya Joseph
Welsh Government	Josh Dixon
Welsh Government	Kathryn Helliwell, Senior Research Officer

Appendix 2: Outstanding questions

1. **How can we work together to improve administrative data 'research readiness'?**
    a. What ways can researchers and/or data owners be more confident about how the data is collected, and use this knowledge to inform their research?
    b. Where does a data custodian's role end in curating and preparing data for use? What is the role of the trusted research environment? How can the work of the researcher be used to iteratively improve the quality of a dataset, and how can this be both acknowledged and rewarded?
2. **How can we best support access to good documentation, and transparency of the audit trail of administrative data processing from collection, through to preparation to address specific research questions?**
    a. What resources and infrastructure would be required for trusted research environments to support access to raw and cleaned versions of datasets, together with well-annotated code documenting these?
    b. What support do data owners need? How can information about how data is collected be encoded and shared?
    c. What would be required for data owners to entrust researchers or third party data processors (such as trusted research environments) to take on the challenge of cleaning the data in its rawest form (with well annotated code?), and making this data as well as cleaned versions available?
3. **How can we best communicate the 'research-readiness' of administrative data in advance of access to the data?**
    a. Are there other models that exist, and/or what are the additional components needed that we could weave into a model such as ABCDE or DUF to make it work best for RR admin data?
    b. Should there be 'minimum standards' for datasets, and what might these be? Should trusted research environments impose these standards?