Office for Statistics Regulation

Regulatory Guidance

Unlocking the value of data through onward sharing

October 2020

Office for Statistics Regulation

We provide independent regulation of all official statistics produced in the UK. Statistics are an essential public asset. We aim to enhance public confidence in the trustworthiness, quality and value of statistics produced by government.

We do this by setting the standards they must meet in the Code of Practice for Statistics. We ensure that producers of government statistics uphold these standards by conducting assessments against the Code. Those which meet the standards are given National Statistics status, indicating that they meet the highest standards of trustworthiness, quality and value. We also report publicly on system-wide issues and on the way statistics are being used, celebrating when the standards are upheld and challenging publicly when they are not.

Unlocking the value of data through onward sharing

"Data is more useful when more people can access and use it. It is most useful when it can be joined together. Data that is inaccessible – or where access takes so long it is rendered irrelevant – is of limited utility." Jeni Tennison, CEO of the Open Data Institute[1]

The central purpose for all official statistics producers is serving the public good through the provision of data and statistics. This obligation is reflected in the principles of the Code of Practice for Statistics which requires statistics producers to commit to, and to promote, the safe onward access to the data used as the basis for producing official statistics. These may include, for example, data from the census, population and business surveys, as well as administrative records.

This guidance is a companion to our guidance on data governance: building confidence in the handling and use of data, which supports data sharing for the public good. It is aimed at Heads of Profession for Statistics and analysts working in producer bodies with an interest in data linkage and sharing.

About this guide

We have written this guidance to increase awareness among statistics producers and users that the principles of the Code of Practice extend beyond statistics production to data sharing and access. We outline practices and processes that uphold these principles. Specific guidance about how to meet these expectations is signposted where available.

We have expectations of producers in three distinct areas:

- Data standards, quality and curation
- Data provision
- Developments to official statistics

Data standards, quality and curation

Documented, consistent, linkable, timely, curated, trackable, reproducible, explorable

Data Provision

Communication and engagement, user-informed processes, transparent processes, appeal and redress, professional development

Developments to official statistics

Improving measures and methods, new insights, improving data quality

[1] https://osr.statisticsauthority.gov.uk/odi-data-blog/

Guidance scope

This guidance applies to:

- providing freely available data as well as more-controlled data (including in safe settings)
- official statistics producers who provide their data via other organisations and those operating their own data access services
- official statistics producers who are also data processors accredited under the Digital Economy Act 2017

Data can be supplied in different formats with restrictions set according to the level of risk of reidentifying data subjects. The corresponding need for added safeguarding measures can also be incorporated, when the steps required will increase in line with the reidentification risk. Completely anonymous open data with no restrictions on users or uses sit at one end of this spectrum. At the other end there is potentially identifiable data that can be accessed only via a terminal on an isolated network in a secure setting, for specific purposes, to approved researchers, with checks made of any analytic outputs. In between these two scenarios there are still restrictions on uses, users and access settings, but these are less extensive. Data owners are responsible for deciding where on this spectrum their data sit, based on the legislation governing their access and the risk environment in which they operate. Sound data governance is underpinned by transparency about such decisions.

Responsibility for data governance and data supply can be split between organisations. For example, the Department for Education in England makes decisions about who can use the data it holds, while access is provided via the ONS Secure Research Service, and direct supply of data to users is permissible only in limited circumstances. The Welsh Government uses the Secure Anonymised Information Linkage (SAIL) Databank to manage access to much of its data. Access to many government surveys is managed by the UK Data Service on departments' behalf.

The DEA created a legal gateway to share de-identified data from public bodies with researchers, to carry out analysis in the public interest, but it specifically does not include data from health and social care services. Before data can be shared for research purposes, it must be processed by an accredited processor so that the data is 'de-identified'. When the data has been de-identified it can be made available to an accredited researcher in a secure environment, and the processor will ensure that any data (or any analysis based on the data) retained by the researcher, or are published, are 'disclosure controlled' to minimise the risk of data subjects being re-identified or other misuses of the data. The accreditation process for researchers, projects and data processors is overseen by the UK Statistics Authority, with decisions taken by an independent Research Accreditation Panel. Access is then provided by one of the accredited processors whose locations span the UK (eight had been accredited as of July 2020). The DEA's Research Code of Practice and Accreditation Criteria sets the standards for this framework.

This regulatory guidance complements the DEA Research Code of Practice. We set out our wider expectations of official statistics producers as providers of open or safeguarded data, some of whom will also be accredited processors under the DEA, but most will not.

The following sections outline our expectations of statistics producers. The Annex contains the specific sections of the Code of Practice for Statistics that producers should pay attention to when reviewing their approaches to data provision and access. We recognise that some of these expectations will be challenging to meet, especially those relating to new areas still undergoing development (for example, reproducibility in safe settings and synthetic data). We have included these more-challenging expectations to highlight areas where statistics producers can work to innovate and improve data provision in the long-term.

Data standards, quality and curation

The full value of data can only be unlocked if they are provided in a format that meets certain standards. Data that are used to produce official statistics should already have been through quality assurance processes, had an assessment of their suitability, and had any known biases identified as part of the original statistics production process. However, for the onward sharing to be effective, data should also be:

- **Documented** and supplied with metadata so that users can understand the provenance, properties and potential of datasets (including linked datasets) and individual data items and can replicate key analyses and variables. Disclosure control steps should also be clearly documented so users can understand how they impact the data, and data subjects can be reassured that their data are being handled safely. Documentation and metadata should be available for users to access before they apply for data unless restrictions are absolutely necessary (for example, for data safeguarding reasons).

- **Consistent** – the data made available should remain consistent with the standards applied at the time the data were collected. Also the data should be consistent with the data used to produce any published estimates – this will require effective version control, good communication between statistics production teams and data supplier teams, and effective mechanisms to update users when changes are made to underlying data after they have been supplied.

- **Linkable** with other data sources – this will require the safe retention of identifiers in standardised formats and a demonstrable public good case for doing so under the General Data Protection Regulations.

- **Timely** – long delays between the publication of statistics and the provision of data diminish the value of the data in the same way that long delays between collection and publication can.

- **Curated** – data increasingly now are curated to support onward use by researchers, rather than just supplied as a by-product of the statistical production process. Data curation is a very clear demonstration of a statistics producer's commitment to increasing the value of data. It involves working with groups of users to identify their needs and developing datasets to meet them. This might involve linking more than one source together, which sometimes requires acting as a broker to work with other data suppliers to source suitable data.

- **Trackable** – enabling the onward tracking of a dataset's use can help to demonstrate the impact of that data and how it has served the public good.

Examples of how this can be done include providing links to published pieces of work in accessible formats, including any lay summaries provided when applications to use the data were originally submitted. Formal mechanisms to track datasets via digital object identifiers (DOIs) are not commonplace at present, but systems are being developed to enable this. Users can also be encouraged to fully cite the data they have used in any published work.

- **Reproducible**[2] – the need for research outputs to be reproducible is recognised as an important way to maintain research integrity. Reproducibility aligns with the Code of Practice's requirements for transparent processes around review and correction, helps to assure quality, and ensures that flawed analyses do not cause harm and undermine the value of data and statistics. To support this principle, data suppliers can place expectations on data users to provide documentation of their data preparation and analysis code. Platforms such as GitHub can help to host code. An archiving policy is also necessary to ensure users can access previous versions for reproducibility purposes. Reproducibility is more challenging when data are provided via secure settings, but models have been developed elsewhere to address this (for example by CASCAD in France).

- **Explorable** – the provisions that need to be in place to safeguard data will often mean that analysts cannot access any data until they have specified their research questions and gained the necessary approvals for their work and themselves. However, sometimes the process of developing research questions requires access to data beforehand to fully understand what can and can't be done. Good metadata and documentation can help here, but it is no substitute for using the data directly. Synthetic data could provide a technical solution here. We encourage data providers to identify ways to support exploratory analyses, including work looking at the potential for synthetic data to achieve this, and at ways to enable an ongoing dialogue with users to support their use.

All these requirements can be supported by data providers working collectively with each other to develop common standards and share best practice. The Working Group for Safe Data Access Professionals, Administrative Data Research UK and Health Data Research UK networks provide these opportunities.

Making data accessible

Data provision comprises two elements:

- decisions about what data access to allow, to whom and for what purpose
- mechanisms for managing the supply of, or access to, the data

There can often be a mismatch between users' and data providers' perceptions of the data access process and how well it meets needs, resulting in frustration on both sides. Many data providers recognise that existing models are not always compatible with how users work and are developing new ways of working to accommodate these. The principles set out in the Code of Practice for Statistics around data provision, and our associated

[2] **Reproducible** refers to the ability to achieve the same research outputs from the same data using the same methods and code. The broader term, **replicable**, refers to the ability to achieve the same research outputs from equivalent or similar data using the same method (but different code).

expectations of what these mean for data providers, provide a framework that could help to better align users and producers' experiences. Data provision that fully supports the standards set out in the Code of Practice for Statistics requires the following:

- **Communication and engagement** with users – our expectations that statistics users will be informed about, and involved in, decisions about official statistics production extend to the provisioning of data. Data providers can hear most easily from their known, existing users – there can be a danger that providers are unaware of the needs of, and barriers limiting access for, potential new users. Users of open data may require more outreach to find. Many complaints that users bring to OSR about data provisioning are a consequence of poor communication about changes to data access arrangements and how data services run. Good communication and engagement involves multiple formats, for example: steering groups or user groups; regular and ad-hoc direct contacts with established networks of users; user surveys; social media; articles and blog posts on relevant websites (including the locations where open data are accessed directly); dissemination via interested networks; organised events with users; attending and presenting at external events that users will be at. Whichever modes are used, having an ongoing dialogue is critical.

- **User-informed processes** – data providers should seek, wherever possible, to ensure that the data provision process reflects how users work and will use the data. Understanding what users need and working to identify solutions that meet those needs are important parts of this process. For example, data users increasingly use automated tools to scrape data, but these won't work if data are provided on sites with CAPTCHA tools, or URLs change without notification or onward direction[3]. There will be times when users' needs cannot be accommodated due to resources, data security considerations or IT infrastructure changes beyond the control of data providers. In these situations, it is important to be transparent about why needs cannot be met. Changes to application processes and data supply mechanisms should be developed in partnership with users to ensure they meet needs and do not have a detrimental impact on their work.

- **Transparent processes** – applying to access data requires proper scrutiny of various elements (for example, project purpose, public good served by using the data, ethics, researcher credentials, IT security) with systems in place to handle each aspect. Transparency about what is needed of applicants and their institutions for each element is essential. Applicants not supplying adequate information is a common reason for applications being rejected and needing to be resubmitted; data providers can help here by ensuring training and support is available to help users navigate the process. Reasons for rejecting applications must be made available to help users understand what more is needed of them, and to uphold the integrity of the decision-making process. Published information about likely timeframes from application submission to data being ready to access is also essential so applicants can plan resources accordingly (especially where staff will be recruited to use the data). This should be based on relevant service performance monitoring data where possible. Once

[3] Examples provided in: Bacon, S. and Goldacre, B. 2020. Barriers to working with NHS England's Open Data. Journal of Medical Internet Research. 22:1. Doi: 10.2196/15603

submitted, users also need access to real-time information about the progress of their application. Clear information about costs for initial applications and any further ongoing costs associated with the data provision must be available.

- **Appeal and redress** – there should be clearly signposted channels for applicants to raise complaints about data provision, including the facility to appeal decisions taken about applications.

- **Professional development** for staff involved in data provision – the professional skills required of the staff who support data provisioning should be recognised and developed in the same way that statisticians who produce statistical outputs are supported to develop their analytical skills. It should also include developing skills in engaging users in an ongoing dialogue.

- **Coordination** across departments in gaining approval – government departments should be working in a coordinated way to make data available to the research community efficiently. Researchers having to get multiple approvals for linked data are exposed to inconsistent decision making and governance standards across different approvals panels – this approach can be inefficient and frustrating for researchers. The legal gateway exists for a coordinated single approvals approach across government, via the DEA, which has robust governance standards approved by Parliament. This approach provides a safe, efficient and consistent gateway for making data available to the research community.

Developments to official statistics

Further analyses of data have long been a source of quality assurance and insight that official statistics producers can draw on. This is reflected in the fact that one of the criteria research projects seeking accreditation under the DEA can use to demonstrate that the work serves the public interest is improving the quality, coverage or presentation of existing research, including official or National Statistics.

For this to work effectively in all settings, not just where the work has this explicit aim, there must be effective communication between data users, data providers and the teams responsible for official statistics production about:

- **Improving measures and methods** – for example new ways of defining or estimating key concepts. This is also important for reproducibility and any accompanying documentation is of wider value to the research community too.

- **New insights** that could usefully become part of routine statistical releases – for example where data have been linked or the analyses has covered topics not included in published outputs.

- **Improving data quality** – particularly where producers can use feedback to help address quality problems at source, or at least before further supply to other users.

Heads of Profession for statistics and lead statisticians should be proactive in ensuring they are aware of projects being undertaken using their data so that these opportunities can be identified.

Useful resources

https://www.ukdataservice.ac.uk/manage-data/document.aspx

https://www.go-fair.org/fair-principles/

https://www.statisticsauthority.gov.uk/about-the-authority/better-useofdata-statistics-and-research/

https://uksa.statisticsauthority.gov.uk/about-the-authority/better-useofdata-statistics-and-research/better-use-of-data-for-statistics/

Annex

What does the Code of Practice say about data sharing and access?

Trustworthiness

T6 Data Governance
Organisations should look after people's information securely and manage data in ways that are consistent with relevant legislation and serve the public good.

T6.1 All statutory obligations governing the collection of data, confidentiality, data sharing, data linking and release should be followed. Relevant nationally- and internationally-endorsed guidelines should be considered as appropriate. Transparent data management arrangements should be established and relevant ethics standards met.

T6.3 Organisations, and those acting on their behalf, should apply best practice in the management of data and data services, including collection, storage, transmission, access, and analysis. Personal information should be kept safe and secure, applying relevant security standards and keeping pace with changing circumstances such as advances in technology.

T6.4 Organisations should be transparent and accountable about the procedures used to protect personal data when preparing the statistics and data, including the choices made in balancing competing interests. Appropriate disclosure control methods should be applied before releasing statistics and data. Appropriate protocols should be applied to approved researchers accessing statistical microdata.

Quality

Q2 Sound methods
Producers of statistics and data should use the best available methods and recognised standards, and be open about their decisions.

Q2.1 Methods and processes should be based on national or international good practice, scientific principles, or established professional consensus.

Q2.2 Statistics, data and metadata should be compiled using recognised standards, classifications and definitions. They should be harmonised to be consistent and coherent with related statistics and data where possible. Users should be provided with reasons for deviations from these standards and explanations of any related implications for use.

Q2.3 Statistics producers should be transparent about methods used, giving the reasons for their selection. The level of detail of the explanation should be proportionate to the complexity of the methods chosen and reflect the needs of different types of users and uses.

Q2.5 Producers of statistics and data should provide users with advance notice about changes to methods, explaining why the changes are being made. A consistent time series should be produced, with back series provided where possible. Users should be made aware of the nature and extent of the change.

Value

V1 Relevance to users

Users of statistics and data should be at the centre of statistical production; their needs should be understood, their views sought and acted on, and their use of statistics supported.

V1.1 Statistics producers should maintain and refresh their understanding of the use and potential use of the statistics and data. They should consider the ways in which the statistics might be used and the nature of the decisions that are or could be informed by them.

V1.2 Statistics producers should use appropriate ways to increase awareness of the statistics and data, communicate effectively with the widest possible audience, and support users and potential users in identifying relevant statistics to meet their needs.

V1.3 User satisfaction with the relevance and usefulness of the statistics and data should be reviewed routinely. This should consider the timeliness, accessibility, clarity and accuracy of the statistics and data.

V2 Accessibility

Statistics and data should be equally available to all, not given to some people before others. They should be published at a sufficient level of detail and remain publicly available.

V2.2 Statistics, data and related guidance should be easily accessible to users. The needs of different types of users and potential users should be considered when determining ways of presenting and releasing the statistics and data.

V2.4 Statistics, data and metadata, including those available through data services, should be released at the greatest level of detail that is practicable to meet user needs. They should be consistent with common data standards and protocols wherever possible.

V3 Clarity and insight

Statistics and data should be presented clearly, explained meaningfully and provide authoritative insights that serve the public good.

V3.1 Statistics, data and explanatory material should be relevant and presented in a clear, unambiguous way that supports and promotes use by all types of users.

V4 Innovation and improvement

Statistics producers should be creative and motivated to improve statistics and data, recognising the potential to harness technological advances for the development of all parts of the production and dissemination process.

V4.1 Statistics producers should keep up to date with developments that can improve statistics and data. They should be transparent in conducting their development activities, and be open about the outcomes and longer-term development plans.

V5 Efficiency and proportionality

Statistics and data should be published in forms that enable their reuse. Producers should use existing data wherever possible and only ask for more where justified.

V5.1 Opportunities for data sharing, data linkage, cross-analysis of sources, and the reuse of data should be taken wherever feasible. Recognised standards, classifications, definitions, and methods should be applied to data wherever possible.