Examples of Data Linking within the Government Statistical Service

This note identifies some further data linking projects employed across the Government Statistical Service (GSS) to complement those mentioned in the National Statistician's response to the Office for Statistics Regulation (OSR) review 'Joining up Data for Better Statistics'. This has a particular focus on examples which demonstrate how some departments are striving for a robust and trustworthy data linking system.

These examples are organised into the six themes presented in the original OSR review to facilitate the development of effective data linkage and statistical production.

It is important to note that the specific data linking processes described will vary depending on the department and the circumstances underlying data-collection and use. Whilst this document does not claim to include a complete account of work being undertaken across the GSS related to data linking, it does intend to provide an overview of how good data linking practice can be developed, employed, shared and be ultimately beneficial to UK research and statistical production.

Contents by Theme; *Joining up Data for Better Statistics*

Theme 1: Government demonstrates trustworthiness, robust safeguarding and clear public communication.

- Actively seek input from the public in major decisions about statistics production and statistical research using data linkage.
- Identify clear, consistent and meaningful language to use when engaging with the public about data safeguarding, linkage and use.
- Be advocates for safe data use to provide insights that serve the public interest.
- Produce "keeping data safe" statements using the Five Safes Framework.
- The Department for Digital, Culture, Media and Sport should develop an overarching framework for trustworthy government data use.

Department for Education (DfE): is aiming to strengthen its transparency and processes around the sharing and linking of data. They routinely publish details of data shares on their website and they are currently working on communications to better explain the protection of confidentiality. They have been working closely with Office for National Statistics (ONS) to make education data available to third parties – usually academics – via the Secure Research Service. This shift means a significant reduction in the extent to which DfE data is being moved, therefore reducing the risk of data breaches and increasing confidence in how the data is being handled.

Office for National Statistics (ONS): require that data acquisition and use is made transparent to heighten the trustworthiness of data use and follow the Code of Practice (CoP) under the Digital Economy Act (DEA).

Scottish Government (SG): have focused its recent communications work on ensuring that primary stakeholders have a high awareness of important 'need to know' data linkage processes, such as accessing data. This is to demonstrate trustworthiness in the sharing and linking of data through robust safeguarding and clear public communications**.** The forthcoming Scottish Data and Informatics Partnership is being established, with effective public engagement identified as a key priority among partners**.** To advocate safe data use, SG require that adequate safe researcher training is completed on the part of those accessing the data, covering legal and ethical approval, proportionate risk assessment and the application of physical and technical safeguards to hosting and provisioning data. Researchers are only able to access linked data through the National Safe Haven providing a controlled and secure environment.

Northern Ireland Statistics and Research Agency (NISRA): have developed a guidance document for its Data Protection Officers and Information Asset Owners outlining the 'Five Safes' network and how to comply with General Data Protection Regulation (GDPR). This has also been shared with NIRSA policy colleagues to share best practice on the safeguarding of data.

Continued:

Theme 1: Government demonstrates trustworthiness, robust safeguarding and clear public communication.

UK Data Service (UKDS): are working to widen access to the data it holds. They recognise that safe data use can provide insight for individuals outside of academia and statistical expertise. They advocate the continued use of the Five Safes Framework, as mirrored through its popularisation of the Administrative Data Research Network (ADRN).

Ministry of Housing, Communities and Local Government (MHCLG): have been successful in linking data for the National Evaluation of the Troubled Families Programme, where data is shared and matched to nationally held administrative datasets every 6 months. The linked administrative data is used to estimate the impact of the programme and to carry out analysis to understand more about factors that affect its outcomes. To advocate safe data use of these data sets, MHCLG have data sharing agreements in place with local authorities and other government departments, that rely on GDPR as the legal basis for processing data and carrying out a task in the public interest.

Department for Digital, Culture, Media and Sport (DCMS): foster links with policy, analytical and scientific colleagues in DCMS, other government departments and external bodies, to maximise strategies for trustworthy data use and answer cross-cutting questions of interest to society.

Theme 2: Data sharing and linkage help to answer society's important questions.

- Maximise opportunities to identify the questions that society wants answered by exploiting existing networks of senior leaders e.g. Heads of Profession, Chief Statisticians, Directors of Analysis and Chief Scientists.
- Ensure that policy makers and external experts are actively involved in processes to identify questions.
- Ensure that departmental Areas of Research Interest Statements explore cross-cutting interests across departments, and are produced with input from all analytical professions.
- Enable more exploratory analysis to take place before research questions are finalised, including through the use of synthetic data.

Home Office: have gained approval from the National Statistician's Data Ethics Committee to carry out experimental work to explore the feasibility of matching Home Office asylum grants to data from the 2011 Census. If successful, this will enable the production of new data on refugee outcomes, providing new policy-relevant insights at a local-level on the socio-economic outcomes of former refugees.

NISRA: is currently developing a Strategic Impact Programme (SIP) as part of the Administrative Data Research Partnership initiative. The development of the SIP has included engagement across a range of groups. NISRA also organised two workshops with senior policy makers and external experts from Northern Ireland universities to identify the key questions that could be addressed through linked

Department for Transport (DfT): have developed a Data Board which pulls together internal and external leads on data in transport to build understanding of work in progress, synergies and potential future developments.

Department for International Trade (DIT): analysts have a programme of academic engagement on trade and investment. DIT have visited and had roundtables with a number of universities across the UK. DIT believes that developing and producing the annual DIT Areas of Research Interest Statements has been, and continues to be, a good focus for engagement.

ONS: is currently running a 'Synthetic data pilot project'. The results from the first phase of work will be published as part of the Methodology Working paper series. The second phase of the project is under-way and focuses on how to synthesise microdata currently held within the Secure Research Service.

HM Revenue and Customs (HMRC): define data linking as a strong probability that an entity is the same on two systems or sources. HMRC usually use this process in operational systems, such as seeking the address of an individual for contact purposes. It largely uses data matching for analytical purposes, such as it's database which was generated to represent the incomes of the population, to help answer societies most important questions.

Theme 3: Data sharing decisions are ethnical timely transparent and proportionate.

- Agree common information governance frameworks to harmonise practice across government departments.
- Consider placing the responsibility for signing off data sharing agreements with more senior staff.
- Explore the contribution that risk assessment tools could make to decision-making about data shares.
- Publish materials related to data shares, including mandatory and voluntary Data Protection Impact Assessments, to support transparency.

Public Health England (PHE): applies anonymisation standards and disclosure control protocols for data releases. PHE has an Office of Data Release (ODR) which determines whether, or how, data can be released. A network of senior data endorsers who are required to sign off users and uses in order for people to access data.

ONS: is in the process of refreshing how it publishes details of the data it holds and the purposes it uses it for. Alongside this, under the transparency require of the Digital Economy Act Code of Practice, information will be published on the intended use for new acquisitions. Data Privacy Impact Assessments are routinely compiled for acquisitions involving personal identifiable data and are shared with data suppliers. ONS routinely produce Data Privacy Impact Assessments to identify risks and mitigations. Independent ethical scrutiny is also provided by the National Statistics Data Ethics Committee where required.

The GSS Data Project: carried out analysis of data products produced across the GSS in 2016. A common approach across the GSS will enable the identification of data sources and capture these for inclusion in future dataset families. When complete, the GSS Data project will provide a registry of data that has a robust discovery metadata framework to ensure a clear line of sight on what information is available and who is responsible for it.

DIT: require microdata access and data sharing agreements to be signed off by the Chief Statistician or Chief Analyst.

SG: is collaborating with data linkage partners and data providers to deliver an improvement programme to change the way decisions are made around the use of data for public benefit research and statistics.

National Records of Scotland (NRS) have a wide representation of stakeholders on governance groups which provide strategic direction and transparent decision-making. This ensures approval processes, procedures and risk assessments are clearly understood and communicated to users of the data, as well as to the public.

Theme 4: Project proposal assessments are robust, efficient and transparent.

- Design data access application processes and support materials with user input and seek ongoing feedback on systems when they are live.
- Signpost users to other sources of data of potential interest to them.
- Work with health data users and NHS Digital to scope the terms of an independent review of NHS Digital's health data sharing and access processes.

ONS: is carrying out a range of user research to ascertain what the areas of difficulty are for users of ONS data, to deepen an understanding of how products and services can be developed and improved to support user needs. This strategy will be continuous and strengthened by the opportunity for users to provide feedback within future prototypes.

The GSS Data Project: seeks to further enhance the data user experience, through an exponentially increasing signpost platform, which can direct users to an ever-growing interconnected data resource on the internet, with informative and interactive properties.

NISRA: highly value feedback from users who access data through secure settings and have recently gauged input from users on their Synthetic Data Policy, Project Modification Policy and Incident Protocol. Changes to application processes are ongoing based off user input.

Department for International Development (DfID): develops all of its data tools through Agile processes, where testing and governance methods mean that user needs are represented in the design of these tools. DfID Data Catalogue enables its staff to crowd source the best sources of data and allows it's colleagues to access this information in a single searchable repository.

SG: is working with partners in UK Statistics Authority (UKSA), ONS and with the other Devolved Administrations through the Research powers in the Digital Economy Act 2017 to deliver a set of assessment arrangements that assure the scientific merit of projects and the training and skills of people working with linked data. This will support the development of a UK-wide system that is coherent and efficient and has simple access points for researchers wishing to conduct diverse research, within, between and beyond the four nations.

NHS Digital: uses a Master Person Service (MPS) to enable data linkage in the wider Data Services Platform (DSP) by producing a Common Base Linkable Attribute (CBLA) that can be used to join National Datasets. This delivers a powerful capability to seamlessly and automatically link the many different aspects of healthcare, including Primary and Secondary Healthcare datasets, and Physical and Mental Healthcare.

Theme 5: Data are documented adequately, quality assessed and continuously improved.

- Identify data sources most in need of additional documentation and metadata and work with data owners to address gaps.
- Create a central registry of administrative data sources.
- Ensure that each step in the data linkage process is documented using a common framework.
- Ensure that statistical experts are consulted and participate in the design process for new data systems.
- Identify mechanisms for data users to feedback information about data quality to data collectors to help improve data at source.

The GSS Data Project: relies on statisticians from across the GSS to feed in their expertise and knowledge of statistics. The project is in regular communications with the GSS Heads of Profession and is building a project board consisting of members of this group. The group engages with statisticians through blogs, conferences and symposiums.

A central repository of administrative sources enables more accurate provenance for the GSS data project and it aims to provide users access to information and offer insight into its origin. The project identifies mechanisms for data users to feedback information about data quality, such as surveys, online remote testing events and group workshops. This feedback is implemented within the project.

ONS: strive to ensure that all documentation for its data assets is complete and consistent, aided through the application of standardised processes to all the data states. For instance, ONS is developing a complete metadata model and will apply it to a new metadata management system built into the Data Access Platform.

To further assist efficient and consistent documentation of data, is the central management of linking core data sets, seen in the ONS' metadata management system or its current work on building analytical pipelines. The ONS consults with statistical experts and works across divisions with individuals from different disciplinary expertise, such as data architects and data suppliers, to enhance the quality of data at its source as well as improving data linking and statistical production.

DfT: is exploring a project to develop a National Access Point (NAP) for all publicly held transport data relating to Roads. In a similar vein to the GSS Data Project, these two areas have now been linked up and a discussion for synergies and potential collaboration is ongoing.

DfID: use it's Data Architecture Programme to systematically map its corporate processes and it's relationship with data systems, to serve as a common language in linking data between systems.

Continued:
Theme 5: Data are documented adequately, quality assessed and continuously improved.

SG: is working with the academic community and with data suppliers to improve the quality of key datasets for research, to ensure these are quality-checked and provisioned with clear meta data. This will involve work to agree standardised metadata tools and methods for data ingestion and will build on expertise within ONS.

PHE: consults academic experts and statistical experts to improve its data linkage methods.

Welsh Government (WG): is developing standards for data dictionaries, metadata and information to provide data sources which have sufficient supporting information. It is also evaluating, with data providers, whether it can provide further supporting information to researchers in line with the Quality Assurance of Administrative Data (QAAD) framework.

NRS is undertaking a range of stakeholder engagement around the census in 2021 which includes considering what other sources of data are available, such as administrative data and linkage opportunities. It aims to improve the clarity pathways for accessing NRS data, balancing the ethical and legal considerations, with the ability for researchers and other stakeholders to access data in a more streamlined and efficient way.

Theme 6: Analysts have the skills and resources needed to carry out high quality data linkage and analysis.

- Recognise resource needs – including the imbalance of demands placed on data holding departments – and either address them with additional inputs, or be clear about what constraints are faced and their implications.
- Government departments to work together to identify resource-efficient solutions to infrastructure requirements (e.g. data storage space, software).
- Identify creative solutions to cut the cost of data extracts charged by external contractors; address this issue in any future contracts.
- Develop a new data linkage skills strategy to support the expansion of opportunities for training and development in this area.
- Identify effective mechanisms to bring in external expertise on data linkage methods and analysis from academics and other experts.
- Ensure the professional development needs of staff who support the data access process are met, including opportunities to network and share practice with people in other organisations in similar roles
- Continue to innovate and share practice around the delivery of safe data settings, especially virtual solutions
- Develop a network of accredited safe settings with common operating standards to act as a single entry point for data users

Department for Business, Energy and Industrial Strategy (BEIS): has delivered a cost effective cloud based specialist analytical IT system (CBAS) to 500 analysts. This provides secure data storage space and allows the department's analysts access to cutting edge software tools for advanced analytics. BEIS has also worked with **DIT** and **DfE** on a resource-efficient solution to share this system. BEIS ensures the professional development needs of staff by providing data science training for all analysts, including sessions in R, QGIS, HMTL and SQL. It also runs a Data Science Project Mentoring Scheme. This gives analysts the opportunity to develop their data science skills by participating in a real project from across the department as part of a small multidisciplinary team.

ONS: is building on the external links already championed in it's data linkage methodology team, so far mainly focused on social and population statistics, and are expanding this network to include external linkage experts in business survey and economic administrative data. Through the continuous development of skills, testing of new methods, defining of best practice and collaboration with existing GSS experts in these topics areas, these hubs will be able to provide training to develop and support a skills strategy across the GSS, and create a community of best practice for continuity. This is further enhanced by GSS Methodology Advisory Committee (MAC) meetings, workshops and tendering research contracts to bring in the best experts across the GSS, UK, international academia and private industry. Similarly, the Analysis Function Strategy (2018) will broaden analytical capability, recognise government data expert from numerous disciplines and establish a core set of skills for non-analysts.

Continued:

Theme 6: Analysts have the skills and resources needed to carry out high quality data linkage and analysis.

Office for Students: use data linkage methods to support business critical processes. It supports the Data Futures Project to create a modernised and efficient approach to collecting data relating to higher education and to deliver better output for a wider range of data users. It has and will continue to develop and share best practice for these methods in conjunction with academics and other government departments.

The GSS Data Project: is seeking to identify resource efficient solutions to demands placed on infrastructure requirements. It is looking to work with reproducible analytical pipeline (RAP) communities to create a common blueprint for RAP processes to feed efficiently into the Linked Data Standards and is currently deliberating federated or centralised strategies to enable this process. Across Government, departments are participating in the RAP champions network, which met for the first time in November.

DfT: has recruited a new central data team of 4 experts to help the department and sector improve the way they approach consistent data standards, data strategy and open data.

DIT: has led a number of trade and investment projects which required a substantial input from academics, leading it to develop programmes for academic engagement, such as it's 'Trade and Economy Panel of Experts'.

Office for Standards in Education, Children's Services and Skills (OFSTED): has focused mainly on data linking plans centred around the ONS Secure Research Service. It has successfully encouraged a number of staff to become approved researchers and to submit project proposals to make use of linked data in this environment. By approaching the delivery of analysis in this way, they enable data to be accessed in a safe and proportionate manner. This also minimises the burden on data holding government departments with requests for data access.