Government Statistical Service

Data for the Public Good

The Government Statistical Service Data Strategy

November 2013

Foreword by the National Statistician

When I launched *Building the Community - The Strategy for the Government Statistical Service to 2020* earlier this year, I said statistics matter and that every day in the UK, important decisions are made based on official statistics. That is still very much the case and underpinning those important statistics is our data. How we use and present our data is critical to ensuring we continue to deliver that service.

Over the past few years, I have seen many changes to the environment in which we operate. Many changes relate to use of our data: from rising demands to measure new and rapidly changing phenomena, to demands for improved delivery times and reductions in response 'burden’, as well as changing user expectations of how they can access, consume and use our data and statistics.

In order to maintain our relevance and increase our impact with data we need to build on our position as leaders in this field. We need to seize opportunities from data, including in the use of administrative data, technology and ‘Big Data’. The vision is of *strengthened use and impact of data for the public good through shared methods, infrastructure and collaboration*. This data strategy is about how we achieve that. It fits closely with a number of cross-government initiatives to increase use and transparency of data, whilst retaining the confidentiality of those that provide it.

This strategy is aimed at the Government Statistical Service, but we need the support of our other professions, our data guardians and information asset owners. To ensure the success of the strategy we need to help them help us. Ensuring our data remains high quality and trusted will be better served through the combined efforts of all our professions.

Through the work with our data that I see from departments, there are huge amounts of innovation, skill and commitment. We must build on this and share our knowledge and successes, and collaborate effectively across the GSS, with our data providers as well as with our data users. We remain in tough financial times and so we must also continue to find ways to be leaner and smarter with our data.

This strategy provides you with the direction and guidance to do this. I encourage you to seek out and innovate in the use of data and data sources and commit to the data strategy vision.

Jil Matheson
National Statistician

1. Executive Summary

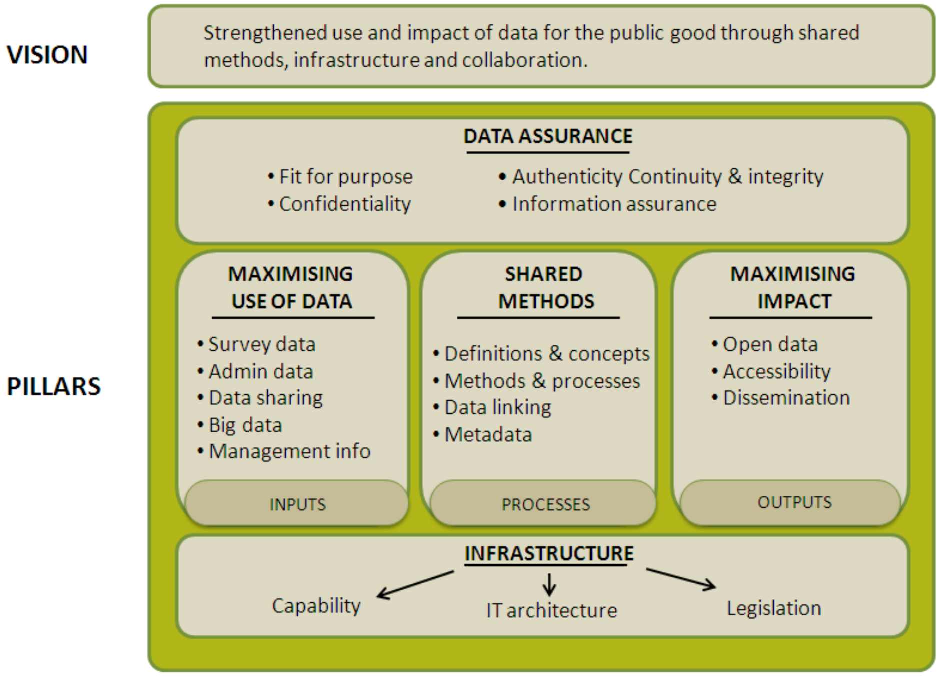
Data are at the heart of the Government Statistical Service (GSS). The GSS specialises in maximising the benefits from data, providing the firm evidence base on which informed decision making is made across government and by the public.

The GSS is experiencing a time of unprecedented change and one of the main drivers of this change is the increasing demand for and value placed on data. The GSS is not alone in this, with official statistics producers internationally facing these same pressures. The key to successfully meeting this challenge is to become flexible in our approach towards data.

The GSS has access to a vast array of high quality data. At the same time the quantity of information available from alternative sources is increasing dramatically and so we must learn to better utilise these sources of data in order to maximise the value this data generates.

Maximising opportunities by working together across the GSS and wider data community will be key to achieving the vision of this strategy. It is clear that we need to make greater use of data (both our own data and others'); making it easily accessible and spending more time helping users to interpret, understand and use it.

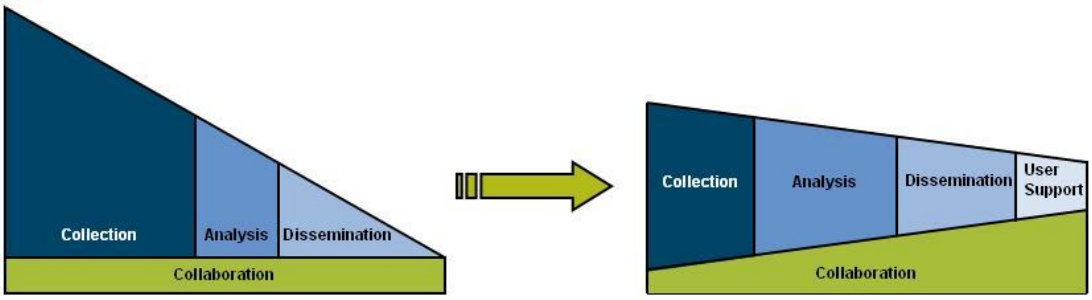
This data strategy provides strategic direction and an overarching plan for the GSS on the data challenges it faces. It takes into consideration the wider perspective of individual departments, devolved administrations and the UK government, as well as their existing and planned activities. The result is a GSS strategic framework for data that not only aims to provide direction for the GSS but to also positively influence the future direction of data initiatives more widely. The GSS framework for data below sets out the vision for data within the GSS.



The vision and making it real

Strengthened use and impact of data for the public good through shared methods, infrastructure and collaboration

The GSS is currently facing two competing pressures. These are the need to make better use of our decreasing resources whilst ensuring that we continue to meet the increasing demand for our services. In order to deliver in this environment, the GSS will have to rebalance its approach to producing statistics. The overarching GSS strategy sets out the need to reduce the focus on data collection through increased collaboration and sharing of data. This will enable us to provide greater analysis and user support.



The GSS can continue to meet the increasing user demand for our data and services by striving towards our vision - strengthened use and impact of data for the public good through shared methods, infrastructure and collaboration. This data strategy aims to do this through focusing on the key areas essential for producing quality statistics, as represented by the five pillars:

- **Infrastructure** – Infrastructure is critical in order for the GSS to effectively deliver the vision for data through innovative IT, working with current legislation and building our people’s expertise in data science.

- **Maximising use of data (inputs)** – Data obtained and used by the GSS reaches maximum value by exploiting administrative and other data sources.

- **Shared methods (processes)** – Data are obtained, used and disseminated by the GSS in a way which makes sharing and reusing easier. This includes harmonised standards, methods, processes and metadata.

- **Maximising the impact of data (outputs)** – Users gain maximum value through increased access and dissemination in ways which meet the needs of different users. This will enable greater informed decision making.

- **Data assurance** – Data are obtained, used and disseminated enabling users to understand their strengths and weaknesses; ensuring new opportunities are balanced against continued confidentiality, security and ethics.

We will maximise the use of new and existing data sources through strengthened partnerships and collaboration. This will mean that the GSS will drive wider economic and social benefits across the UK through the value that our expertise in data will add.

2. Infrastructure

The right skills, technology and legal framework in place in order to deliver the vision for data within the GSS.

Our infrastructure enables us to best utilise and add value to data. Its continual development is essential, underpinning our ability to achieve our vision for data.

The GSS needs to become digital by default, maximising the use of existing technologies whilst ensuring it makes the most of those that are emerging. We will work more closely within the GSS, across government, and the wider community to do this. We will use innovative IT to develop shared solutions and build expertise in data science. We will work with current and emerging legislation and fully exploit alternative ways of making data available where the legal framework limits how we can share data.

IT architecture

New technologies are transforming the way data are captured, stored, processed and disseminated. The GSS will be at the forefront of embracing new innovations that enhance its activities and seek to adopt those which improve its efficiency and effectiveness.

We will work together to achieve better value by reusing and sharing ICT solutions. Where possible, we will make developments interoperable between departments and apply appropriate common standards so that data across the GSS are compatible. This will help data to reach its maximum utility. It will be essential to build our relationships with the ICT community in order to achieve this.

Legislation

Legislation, or the lack of specific legislation, is perceived to be one of the biggest current barriers to sharing data across government. Under the current situation only limited data sharing can occur and so it can be difficult to realise the benefits and maximise the potential.

We will make the best use of the current legislation and engage proactively with any development of data sharing legislation. In addition, we will also work to change the culture and attitudes within government on data sharing. Legislation in itself should not be a barrier to sharing data and the GSS will continue to actively pursue other avenues for making data available, including open data.

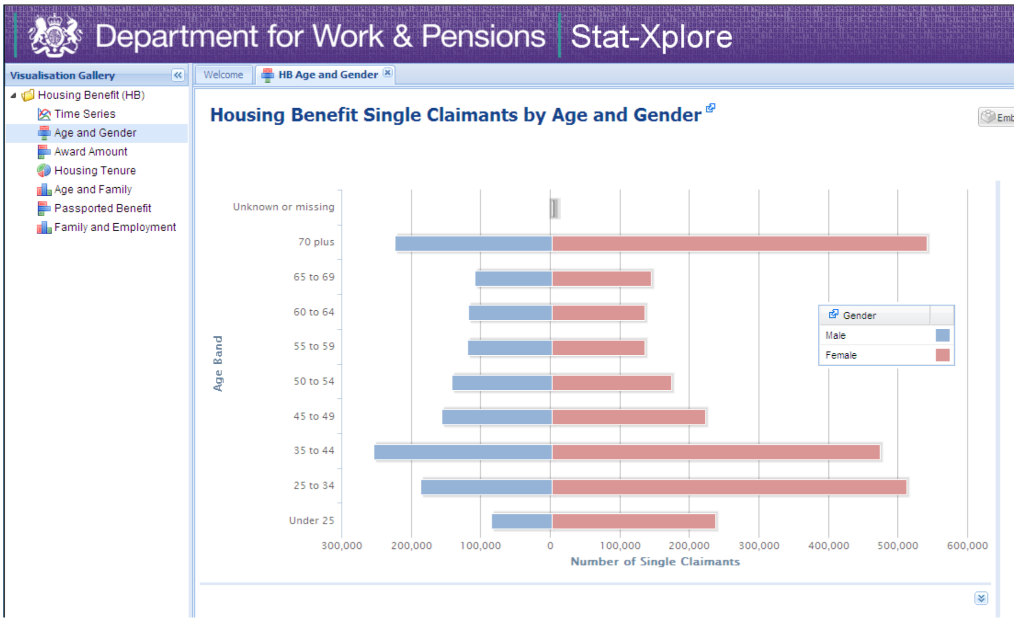
Capability

The GSS encourages and supports investment in professional development and sharing of knowledge to underpin, support and enhance best practice for government and the broader public good.

The vast opportunities presented by new technologies will require new expertise and new relationships. We will continue to build on our current expertise, and identify and enhance our skills that contribute to emerging fields such as data science. We will collaborate more closely: across the GSS, with professions such as ICT, with academia and commercial organisations. This will provide the GSS with the capability and capacity to continue to embrace new data challenges and opportunities.

Case Study: DWP Stat-Xplore

DWP Stat-Xplore was developed as through collaboration between statisticians and commercial software and web hosting companies. It is an innovative web product allowing users access to the utility of microdata, whilst keeping the user separate from the data itself. It gives users the flexibility to interrogate their data in an intuitive way, producing user friendly outputs. These outputs use consistent disclosure control to ensure that legislative requirements on protecting confidential information are fulfilled.



3. Maximising the Use of Data

Data obtained and used by the GSS reaches its maximum value.

Our vision is to collect data once, use many times and to ensure we achieve maximum value and impact.

We are privileged to already have access to a wealth of data. However, we need to do more to make the best use of all types of data, whatever its source. We will take opportunities to be more innovative by exploiting existing data sources and exploring opportunities to access and use new data. We will work collaboratively to achieve this to better support decision making and to better meet user needs.

There are significant benefits to maximising the use of a range of data sources. That can be as simple as taking time to better understand the quality and relevance of our data sources, or improving the evidence base by investigating more data linking and data sharing within and beyond the GSS.

Better use of existing data

Increased and more effective use of administrative data has the potential to reduce our reliance on survey collections, which will enable us to minimise burden and costs. However, administrative data (or management information) is not collected for statistical purposes, so we need to develop systems and processes for rigorously auditing its quality and better understanding how it can be used for statistics.

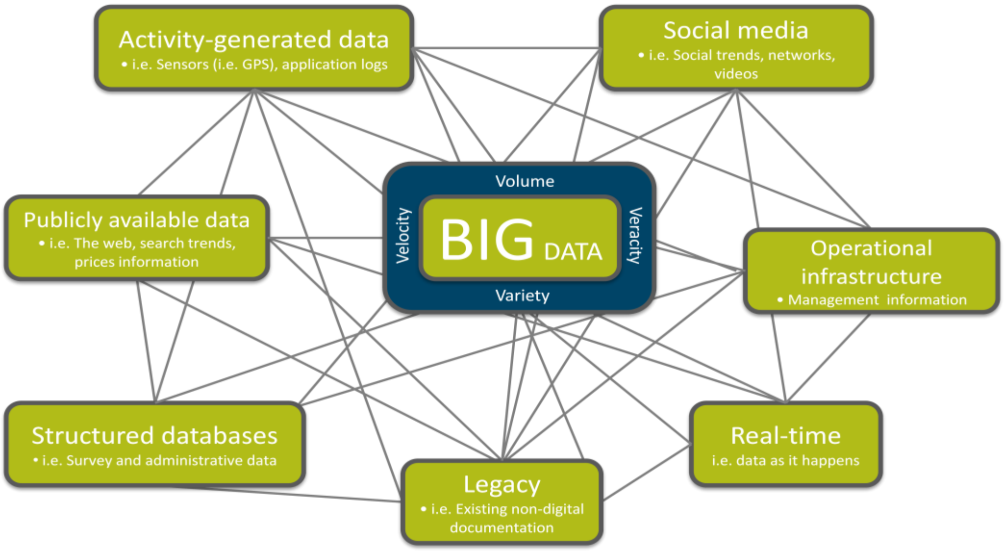
- **Data sharing –** We will prioritise data sharing initiatives, increasing our capability and capacity for this where it is ethical and legal to do so. By working together across government and with our users, we will identify opportunities for setting up new data shares, and continue to share our experience of best practice in this field. Data sharing will enhance the evidence base in an efficient way and help to solve complex issues that cannot be answered by dealing with data sets in isolation.

New data sources

In order to ensure our data and statistics are value for money and relevant, we need to be creative and continually explore whether there are new or alternative data sources that could be used. The GSS will explore the potential, risks and opportunities of using these emerging data sources to meet user needs.

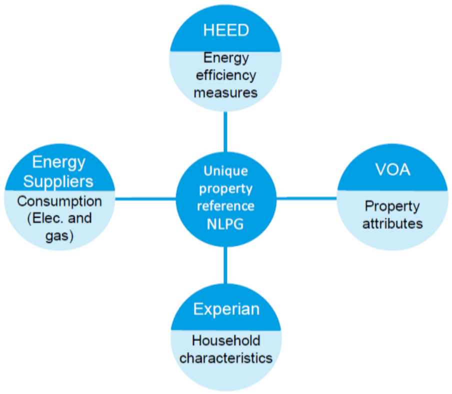
- **Big Data** – The term Big Data is used to describe data which are high in volume, velocity, veracity, and variety. In reality this term is often used quite broadly to mean a variety of data sources (Figure 1). The GSS already have experience in types of Big Data including structured databases, operational infrastructure and real time systems. We will continue to build on this and look for new opportunities where big data could potentially supplement our existing data sources.

Figure 1. The Big data landscape.


Activity-generated data
• i.e. Sensors (i.e. GPS), application logs
Social media
• i.e. Social trends, networks, videos
Publicly available data
• i.e. The web, search trends, prices information
Volume
Velocity
BIG DATA
Veracity
Variety
Operational infrastructure
• Management information
Structured databases
• i.e. Survey and administrative data
Legacy
• i.e. Existing non-digital documentation
Real-time
i.e. data as it happens

Case Study: The National Energy Efficiency Data Framework

The National Energy Efficiency Data Framework (NEED) was set up by DECC to provide a better understanding of energy use and energy efficiency in domestic and non-domestic buildings in Great Britain. The data framework matches gas and electricity consumption data collected for DECC sub-national energy consumption statistics and records of energy efficiency measures from the Homes Energy Efficiency Database (HEED) run by the Energy Saving Trust (EST). It also includes data about dwellings and households obtained from a variety of sources. Data in NEED currently covers the domestic and non-domestic sectors across the whole of Great Britain. By bringing together these different data sources, DECC was able to obtain greater insight into energy usage than if it the individual data sources had been used in isolation.


HEED
Energy efficiency measures
Energy Suppliers
Consumption (Elec. and gas)
Unique property reference NLPG
VOA
Property attributes
Experian
Household characteristics

Shared Methods

Data are obtained, used, and disseminated by the GSS through harmonised standards, methods and processes.

One of the greatest challenges to ensuring data are fully exploited is ensuring compatible classifications, definitions and processes are adopted. These enable data from different sources to be used together, significantly expanding their value and utility. We will extend the use of harmonised standards, methods and processes, together with more comprehensive and increased use of metadata

Definitions and concepts

Government data are many and varied, designed at different times by a range of departments to meet specific needs and purposes. Although there is no "one size fits all" solution, we can do more to harmonise data definitions and concepts by encouraging wider adoption of standard classifications and concepts.

Methods and processes

Both the production of our data, and the analyses and statistics that come from it need to follow consistent transparent processes in order to ensure repeatable results. This applies from data collection through to processing, analysis and dissemination.

- **Data linking and matching** – Data linking and matching is a key enabler to this data strategy. Therefore issues around consistent processes are particularly key here. Such 'joins' between data need to be repeatable and of a known quality in order to ensure consistency of the results.

Metadata

Metadata are critical as they define and describe the data. Without metadata there is no clear understanding of how data can be used and how they can be best exploited. We will make sure that comprehensive metadata, whether contextual or technical, are available in compatible formats, openly and consistently. This will make it easier for users of the data to reuse it in appropriate ways.

Case Study: INSPIRE

European Commission

INSPIRE

Infrastructure for Spatial Information in the European Community

The INSPIRE directive came into force in 2007 and aims to create a European Union spatial data infrastructure. This will enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe.

A European Spatial Data Infrastructure will assist in policy-making across boundaries. Therefore the spatial information considered under the directive is extensive and includes a great variety of topical and technical themes.

INSPIRE is based on a number of common principles:

- Data should be collected only once and kept where it can be maintained most effectively.
- It should be possible to combine seamless spatial information from different sources across Europe and share it with many users and applications.
- It should be possible for information collected at one level/scale to be shared with all levels/scales; detailed for thorough investigations, general for strategic purposes.
- Geographic information needed for good governance at all levels should be readily and transparently available.
- Easy to find what geographic information is available, how it can be used to meet a particular need, and under which conditions it can be acquired and used.

Maximising the Impact of Data

Users gain maximum value from our data in order to support informed decision making.

Users are the reason why we produce our data and statistics. In order to help them get the most out of our data we will ensure data are easily discoverable, accessible, understandable and usable.

Accessibility

Improved user access to our data for researchers and the public will result in more and improved use of our data. An essential part of improving user access is being able to help our users find out about our data in the first place, where to find it and know how to use it once they do have access.

- **Open data** – The GSS has been a long standing supporter of open data. The open data approach allows our data to be used together, whatever the source or location; allowing for the provision of applications which can act as a single access point if required. The GSS will continue to make its data increasingly open and accessible, unlocking value and potential.

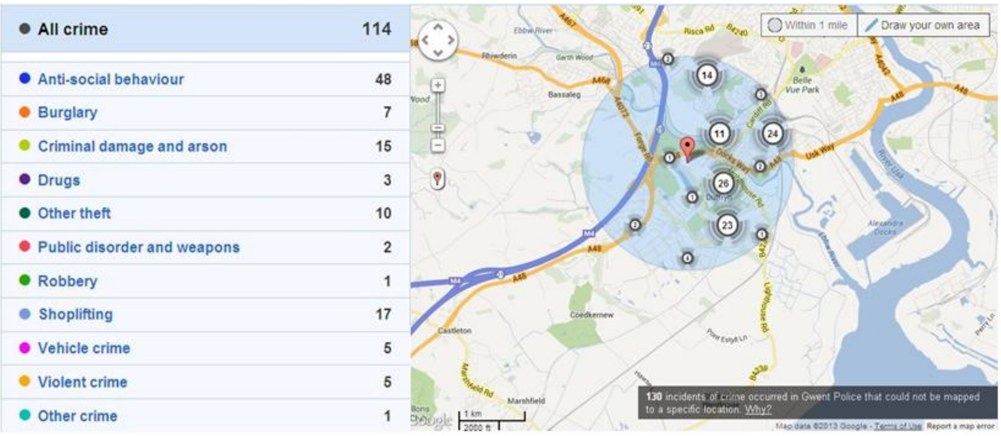
Dissemination

There are already a number of existing ways our statistics are disseminated, both physically and online. In order to meet increasing user demand for our data we must make better use of tools that allow users (depending on who they are) to gain most value whilst ensuring confidentiality. This includes:

- **Web-based data** – For aggregate data which underlies our statistics it will be key to not only place data on the web (i.e. as a CSV file) but actually in the web so that it can be directly machine read. This will allow freedom to reuse the data underlying our statistics.
- **Remote job submission systems** (microdata utility without direct microdata access) – An increasing amount of users are requiring the use of microdata. In order to meet this need effectively and securely, systems can be put in place where users gain the utility of microdata without having direct access. Stat-Xplore, mentioned early in this strategy, is an example of this.
- **Data research centres** – Where direct access to government microdata is the only option, data research centres should be used in order to provide access to this data in a safe environment.

Case Study: POLICE.UK

POLICE.UK – The Home Office and Ministry of Justice have made crime data available which can be searched by postcode. Information such as number of crimes reported each month, outcomes and types are available. This transparency allows citizens to form a clearer idea of their surrounding area and make choices (such as moving home, or planning a route home at night) accordingly. Underlying data are available through the website and are also accessible by an API[1].



● All crime	114
● Anti-social behaviour	48
● Burglary	7
● Criminal damage and arson	15
● Drugs	3
● Other theft	10
● Public disorder and weapons	2
● Robbery	1
● Shoplifting	17
● Vehicle crime	5
● Violent crime	5
● Other crime	1

130 incidents of crime occurred in Gwent Police that could not be mapped to a specific location. Why?

[1] An Application Programming Interface (API) is a way of allowing computers direct access to data on the web without the need for any human interaction, This enables data to be openly available for reuse.

Data Assurance

Data are obtained, used and disseminated in a way that ensures continued trust by suppliers of data and its users.

Maintaining trust in official statistics is essential. Trust matters because it affects the utility of our statistics, both in terms of the value to society and in terms of ensuring continued access to data to construct statistics in the first place. Central to maintaining trust is the assurance that the data we use are fit for purpose and that we can continue to protect that data.

Fit for purpose

With a move towards more data sharing (collecting data once and using many times), data are more likely to be used for purposes that go beyond what it was originally collected for. It is therefore even more important to ensure that the source dataset, underlying systems and metadata are well documented. We need to ensure that statistical staff develop an awareness of the multiple purposes of data and have the tools and skills to assess and communicate data quality. This will enable users to make informed decisions on whether they choose to use the data for their purpose.

Confidentiality

We have a duty to protect data that we collect from our citizens and businesses, and to ensure that they are suitably informed of our likely use of this data.

As data are increasingly made available and linkable through common architecture, the job of protecting confidentiality becomes increasingly complex. We will continue to adapt the way we deal with confidentiality. We will ensure that our approach recognises the public’s demands for more data whilst adequately protecting the data providers’ anonymity.

Information Assurance

We have a legal and ethical duty to ensure that personal and commercial data are kept safe – to the level of security appropriate to the data.

As more use is made of cloud-based IT solutions and other technology we will continue to keep data safe by having the right systems, guidance and training in place.

- **Authenticity, Integrity and Continuity** - As more administrative and other data are used to produce statistics we must pay particular attention to ensuring that the data continues to be an authentic representation of the subject matter;

ensuring it continues to be used ethically; and ensuring that the impact on supply and established statistical series are minimised.

Case Study: HMRC VAT

HMRC VAT – Whilst VAT data is critical for HMRC's own functions, it is also supplied to ONS for its Interdepartmental Business Register (IDBR), which along with quarterly PAYE scheme data feed, allows ONS to create enterprises which are the business entities on the IDBR. This is record-level data, shared through the Value Added Tax Act 1994 for VAT traders and the Finance Act 1969 for PAYE employers. The IDBR has multiple uses and users from central and local government, MPs and private sector to academia and regularly by the media who report on business start-ups and closures as an indicator on economic activity. The products are generally used by those who wish to examine the numbers of businesses in certain industries and or geographical areas.

Extracts of HMRC VAT data are also placed into the HMRC Datalab and the available variables are documented online: http://www.hmrc.gov.uk/datalab/vat-dataset.htm. The HMRC Datalab was launched in May 2011 as a new Research Data Centre (RDC). It allows approved academics to access anonymised HMRC data in a secure environment that is consistent with the department's data security policy. The aim of this initiative is to produce high quality research that benefits both the department and the wider academic community, in the form of a wider evidence base to support knowledge sharing and policy making.

Annex A – Implementation and Delivery

Implementation is key to ensuring the GSS is able to successfully deliver the data strategy. As such, the key actions required to successfully deliver the strategy, our vision and key outcomes are set out below. Further detail on these actions, together with the projects required to achieve them, their responsible owners and timeframes are set out in the accompanying implementation plan.

Pillar	Outcome	Actions
Infrastructure	The right skills, technology and legal framework in place in order to deliver the vision for data within the GSS.	• Ensure the GSS plays a key role in data legislation and other data initiatives across government • Ensure data skills become a core requirement for GSS members' career development. • Ensure the GSS encourages more cross-specialism experts within government. • Invest time using innovative IT to exploit our data.
Maximising the use of data	Data obtained and used by the GSS reaches its maximum value.	• Create a greater shared understanding of our data. • Regularly evaluate data sources for potential new uses. • Increase utility of new data sources. • Work with data providers to improve the utility of existing data sources.
Shared methods	Data are obtained, used, and disseminated by the GSS through harmonised standards, methods and processes.	• Promote standardised classifications of data. • Data and their metadata are available together in compatible formats; openly and consistently. • Shared methodology for data processing will be encouraged across the GSS.
Maximising the impact of data	Users gain maximum value from our data in order to support informed decision making.	• React positively and identify opportunities for our data from users' data needs and their outputs from our data. • Share our data better, in clearer and wider ways.
Data assurance	Data are obtained, used and disseminated in a way that ensures continued trust by suppliers of data and its users.	• Ensure quality is maintained whilst enabling our data to be fit for multiple purposes • Continue to keep data safe, maintaining and building trust.

Government
Statistical Service